Statistical Methods for Multivariate and Complex Phenotypes

Statistical Methods for Multivariate and Complex Phenotypes

A dissertation presented

by

Denis Madison Agniel

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

August 2014

Dissertation Advisor: Professor Tianxi Cai Denis Madison Agniel

Statistical Methods for Multivariate and Complex Phenotypes

Abstract

Many important scientific questions can not be studied properly using a single measurement as a response. For example, many phenotypes of interest in recent clinical research may be difficult to characterize due to their inherent complexity. It may be difficult to determine the presence or absence of disease based on a single measurement, or even a few measurements, or the phenotype may only be defined based on a series of symptoms. Similarly, a set of related phenotypes or measurements may be studied together in order to detect a shared etiology. In this work, we propose methods for studying complex phenotypes of these types, where the phenotype may be characterized either longitudinally or by a diverse set of continuous, discrete, or not fully observed components. In chapter 1, we seek to identify predictors that are related to multiple components of diverse outcomes. We take up specifically the question of identifying a multiple regulator, where we seek a genetic marker that is associated with multiple biomarkers for autoimmune disease. To do this, we propose sparse multiple regulation testing (SMRT) both to estimate the relationship between a set of predictors and diverse outcomes and to provide a testing framework in which to identify which predictors are associated with multiple elements of the outcomes, while controlling error rates. In chapter 2, we seek to identify risk profiles or risk scores for diverse outcomes, where a risk profile is a linear combination of predictors. The risk profiles will be chosen to be highly correlated to latent traits underlying the outcomes. To do this, we propose semiparametric canonical correlation analysis (sCCA), an updated version of the classical canonical correlation analysis. In chapter 3, the scientific question of interest pertains directly to the progression of disease over time. We provide a testing framework in which to detect the association between a set of genetic markers and the progression of disease in the context of a GWAS. To test for this association while allowing for highly nonlinear longitudinal progression of disease, we propose functional principal variance component (FPVC) testing.

Contents

Title page i

Abstract iii

Table of Contents iv

1 Identifying predictors for multiple outcomes using semiparametric models **1**

1.1 Introduction 2

1.2 Stepdown testing 5

1.2.1 Testing a single predictor x_j 6

1.2.2 Controlling error rates across all predictors 8

1.3 Estimation 8

1.3.1 Asymptotic Theory 9

1.3.2 Estimating the variability in $\widehat{\boldsymbol{\beta}}$ 10

1.4 Tuning 11

1.4.1 Choosing a reference distribution 12

1.4.2 Choosing ψ 13

1.4.3 Choosing λ 14

1.5 Example 14

1.5.1 Genetic study to identify shared autoimmune risk loci 14

1.5.2 Simulation results 16

1.5.3 Bias, standard errors (SEs), and confidence intervals (CIs) 18

1.5.4 Testing 18

1.6 Discussion 21

1.7 Appendix . . . 23
1.7.1 Justification of stepdown procedure . . . 23
1.7.2 Proof of sparsistency and asymptotic normality . . . 24
1.7.3 Properties of resampled $\widehat{\boldsymbol{\beta}}^*$ . . . 26
1.7.4 Algorithm . . . 27

2 Semiparametric canonical correlation analysis 29
2.1 Introduction . . . 30
2.2 Semiparametric canonical correlation analysis . . . 32
2.2.1 Identifying $\mathbf{h}$ . . . 33
2.2.2 Estimating the joint distribution . . . 34
2.2.3 Assessing variability . . . 36
2.2.4 Visualizing risk profiles . . . 37
2.3 Example . . . 37
2.3.1 Genetic study to identify risk profiles for autoimmune disease . . . 37
2.3.2 Simulation results . . . 41
2.4 Discussion . . . 47
2.5 Appendix . . . 52
2.5.1 Expansions of $\widehat{\boldsymbol{\sigma}}$ . . . 52

3 Genome-wide association studies of longitudinal outcomes 55
3.1 Introduction . . . 56
3.2 Functional principal variance component testing . . . 58
3.2.1 The test statistic . . . 58
3.2.2 Connection to mixed effects models . . . 62
3.2.3 Estimating the null distribution of the test statistic . . . 64
3.2.4 Combining multiple sources of outcome information . . . 65
3.3 Simulation results . . . 66
3.4 Association between genetics and trajectory of LDL . . . 71
3.5 Discussion . . . 75

3.6 Appendix 76
3.6.1 FPCA Assumptions 76
3.6.2 Justification for the asymptotic null distribution 77
3.6.3 Justification for the form of the test statistic 80

Identifying predictors for multiple outcomes using semiparametric models

Denis Agniel, Katherine P. Liao, and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

1.1 Introduction

In recent years, considerable interest has been focused on studying multiple phenotypes simultaneously in both epidemiological and genomic studies. There are several reasons for such studies to be important. First, a complex disorder is usually associated with multiple correlated phenotypes. Hence, even when the focus of the study is on a single disease, multiple phenotypes might be needed to fully capture the complexity and multidimensionality of the disorder. Second, multiple related disorders might share the same etiology and a joint assessment will enable researchers to identify factors associated with risk of multiple diseases. In genetics, researchers might hypothesize that a group of related diseases share a common genetic basis. As an example, recent studies have identified common genes associated with a higher risk of what were previously considered distinct autoimmune diseases (Zhernakova et al., 2009; Xavier and Rioux, 2008). Similar shared genetic bases have also been suggested for various types of cancers and related psychiatric disorders (Solovieff et al., 2013). Identification of predictors of multiple outcomes, also commonly known as multiple *traits* in the genetics literature, can improve understanding of disease etiology, genetic regulatory pathways, and treatment. Further complicating matters, the measurements of the outcomes may be *diverse*: they may be binary (e.g., presence of disease), continuous (disease activity score), ordinal (severity of disease), not completely observable (perhaps due to a limit of quantification), or any combination thereof.

To address these questions statistically, we seek to assess the association between a vector of predictors $\mathbf{x} = (x_1, ..., x_p)^\intercal$ and a vector of outcomes $\mathbf{y} = (y^{(1)}, ..., y^{(M)})^\intercal$ by estimating and testing all relevant effects. For each predictor x_j we desire an estimation and testing procedure that will identify its associated subset of $\mathbf{y}$. In particular, researchers often want to identify predictors that are important for multiple or all outcomes. We will call x_j a "multiple regulator" if it is associated with more than one outcome, a terminology which we adapt from Peng et al. (2010). An example of what we call multiple regulation is known as pleiotropy in the genetics literature. Our goal of identifying multiple regulation is not to be confused with identifying predictors that are associated with any outcomes. Association with any outcomes has been an active area of research in recent years, with two examples being

global association tests and group-sparse regularization. Association tests provide a test for the relationship between x_j and the entire set $\mathbf{y}$ (Lange et al., 2003; Jiang and Zeng, 1995) and have been shown in some situations to have higher power than marginal tests to detect associations when x_j relates to multiple outcomes (Zhu and Zhang, 2009). Group-sparse methods, largely based on the group lasso (Yuan and Lin, 2006), select predictors that are relevant for any outcome (Turlach et al., 2005; Obozinski et al., 2011).

Here, we are particularly interested in identifying predictors that are relevant for multiple outcomes and inferring which subset of $\mathbf{y}$ each of the x_j's are associated with. There is a paucity of literature that addresses these specific questions. Under linear regression models, the remMap procedure (Peng et al., 2010) addresses such a question via variable selection by jointly penalizing both the L_1 and L_2 group norms of a squared loss. Under generalized linear models, one could potentially modify the hierarchical lasso (Zhou and Zhu, 2010) procedure, originally proposed to handle grouped predictors with a single outcome, to address the multiple regulator problem. However, these methods are not applicable when $\mathbf{y}$ consists of a mixture of outcomes with different scales.

Furthermore, regardless of estimation technique, a multiple testing procedure is required to control error rates when identifying multiple regulation, which operates on the (potentially large) set of hypotheses $\{H_j^{(m)} : x_j \text{ unassociated with } y^{(m)}\}_{j=1,\ldots,p;m=1,\ldots,M}$. None of the existing methods for multiple regulation tackles this issue. In general, multiple testing based on regularized estimation is challenging for two reasons. First, while many of the regularization procedures such as Zhou and Zhu (2010) established asymptotic *oracle properties* for their estimators — non-informative predictors can be detected with no uncertainty and their detection induces no additional variation in the estimation of the informative predictors (Fan and Li, 2001; Zou, 2006) — in finite samples those properties may be far from holding. Consequently, basing testing procedures on such asymptotic results may lead to inflated type I error in finite samples. Second, the estimators and hence their corresponding test statistics could be highly correlated from the regression fitting. Standard methods, including the Bonferroni procedure to control the familywise error rate (FWER) and various false discovery rate (FDR) controlling procedures (Benjamini and Hochberg, 1995), ignore

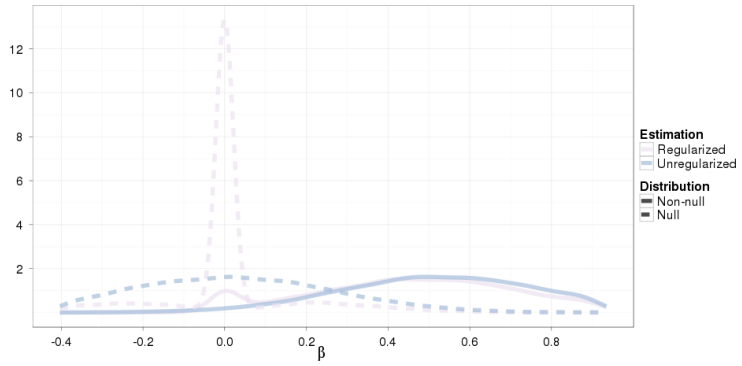


Figure 1.1: Sampling distributions of null and non-null effects, with and without regularization. Tails of the distributions are truncated for ease of presentation.

information about the joint distribution of the test statistics outside of the ordered p-values and either require positive dependency or tend to be conservative.

We propose a two-stage technique to both estimate the effects of $\mathbf{x}$ on $\mathbf{y}$ and identify multiple regulation while controlling error rates. In the first stage, we use regularization to induce sparsity in the estimated effects. To do this, we generalize the adaptive hierarchical lasso of Zhou and Zhu (2010) to handle the case of semiparametric models, which is necessary because $\mathbf{y}$ may contain components that are not fully observed. In the second stage, we employ a stepdown procedure analogous to Romano and Wolf (2005) to identify multiple regulation while controlling error rates. Our two-stage testing method, entitled Sparse Multiple Regulation Testing (SMRT), is powerful for several reasons. First, regularization enables us to more efficiently estimate both the null and non-null effects. The null effects are estimated as 0 with probability tending to 1 and the non-null effects are estimated with lower variability compared to unregularized estimators. Furthermore, the distributions of the estimates of null effects, which tend to a point mass at 0, and the distributions of the estimates of non-null effects, which tend to exclude 0, are distinctly separated through regularization, giving us more power to detect the non-null effects (see figure 1.1 for an illustration from our simulations). However, it is generally challenging to perform testing based on regularized estimators since their distributions in finite samples cannot be approximated well by asymptotic results. We lay out permutation- and resampling-based procedures to better

approximate the distributions of the proposed test statistics and the regression parameter estimators. This enables us to properly control error rates for both hypothesis testing and interval estimation.

The rest of the paper is organized as follows. In section 1.2, we describe the testing procedure. In section 1.3, we discuss estimation using the hierarchical lasso for sparse semiparametric regression with multiple outcomes, we provide asymptotic properties of the estimator, and we introduce a method to estimate its variability. In section 1.4, we discuss issues related to tuning for both estimation and testing. In section 1.5, we apply our method to a genetic study of autoantibodies with the goal of identifying multiple regulators of autoimmunity. We also provide simulation results which validate our method. And finally, in section 1.6, we discuss implications and further directions of our method.

1.2 Stepdown testing

Suppose the data for analysis consists of n independent and identically distributed random vectors $\mathbb{V} = \{\mathbf{V}_i = (\mathbf{y}_i^\intercal, \mathbf{x}_i^\intercal)^\intercal\}_{i=1,...,n}$ where $\mathbf{y}_i = (y_i^{(1)}, ..., y_i^{(M)})^\intercal$ are the M outcomes and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^\intercal$ are the p predictors for the the ith subject. We assume that

$$P(y^{(m)} \leq y \mid \mathbf{x}) = G^{(m)}\{\mathbf{x}^\intercal \boldsymbol{\beta}_0^{(m)} + h^{(m)}(y)\} \qquad m = 1, ..., M, \tag{1.1}$$

where $\boldsymbol{\beta}_0^{(m)}$ represents the unknown effect of $\mathbf{x}$ on $y^{(m)}$, $h^{(m)}(\cdot)$ is an unspecified smooth, increasing function, and the link function, $G^{(m)}$, is given although the correlation structure of $\mathbf{y}$ is left unspecified. For ease of presentation, we assume that $\mathbf{y}$ is fully observed although the proposed method can easily accommodate censored outcomes. When $y^{(m)}$ is continuous, (1.1) is equivalent to

$$h^{(m)}(y^{(m)}) = -\mathbf{x}^\intercal \boldsymbol{\beta}_0^{(m)} + \epsilon^{(m)}, \quad \epsilon^{(m)} \sim G^{(m)}. \tag{1.2}$$

For binary or ordinal outcomes, (1.1) and (1.2) are only defined at certain threshold values and correspond to parametric models. Choice of $G^{(m)}$ determines the type of model being fit. For example, $G^{(m)}(x) = e^x/(1+e^x)$ corresponds to a proportional odds model for continuous $y^{(m)}$ and a logistic regression model if $y^{(m)}$ is binary. One may let $G^{(m)}(x) = 1 - e^{-e^x}$ to impose a proportional hazards model.

To estimate $\boldsymbol{\beta}_0^{(m)}$, one may employ the non-parametric maximum likelihood estimator (NPMLE) under model (1.1) (Zeng and Lin, 2007; Murphy et al., 1997; Murphy and Van der Vaart, 2000) – or the maximum likelihood estimator if (1.1) corresponds to a parametric model – based on data observed on the mth outcome, $\mathbb{V}^{(m)} = \{(y_i^{(m)}, \mathbf{x}_i^\intercal)^\intercal\}_{i=1,...,n}$. Let $\mathcal{L}^{(m)}(\boldsymbol{\beta})$ denote the resulting profile log-likelihood (PLL) function corresponding to the NPMLE. It has been shown that under mild smoothness conditions, the profile likelihood can be treated as a regular likelihood, and the maximum PLL estimator $\widetilde{\boldsymbol{\beta}}^{(m)} = \text{argmax}_{\boldsymbol{\beta}}\, \mathcal{L}^{(m)}(\boldsymbol{\beta})$ is regular and semiparametric efficient (Murphy and Van der Vaart, 2000). However, when p is not too small and $\{\boldsymbol{\beta}_0^{(m)}\}_{m=1,...,M}$ might be sparse, an improved estimator may be obtained by imposing regularization on the PLL. Let $\widehat{\boldsymbol{\beta}}$ denote the regularized estimator of $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{(1)^\intercal}, ..., \boldsymbol{\beta}_0^{(M)^\intercal})^\intercal$ which we detail in section 1.3. To develop a testing procedure based on $\widehat{\boldsymbol{\beta}}$, we employ a reference distribution (discussed further in section 1.4) to estimate the null distribution of $\widehat{\boldsymbol{\beta}}$.

1.2.1 Testing a single predictor x_j

In order to make inference on a single predictor, SMRT employs a stepdown procedure for x_j considering the M hypotheses $\mathcal{H}_j = \{H_j^{(m)} : \beta_{0j}^{(m)} = 0\}_{m=1,...,M}$ with alternative hypotheses denoted $\{\bar{H}_j^{(m)} : \beta_{0j}^{(m)} \neq 0\}_{m=1,...,M}$.

To test $H_j^{(m)}$, we consider the test statistic $t_j^{(m)} = n^{\frac{1}{2}} \left|\widehat{\beta}_j^{(m)}\right| / \widetilde{\sigma}_j^{(m)}$ and its reference distribution $\mathcal{T}_j^{(m)} = \{t_j^{*b(m)}\}_{b=1,...,B}$which approximates the distribution of $t_j^{(m)}$ under $H_j^{(m)}$ and can be obtained by resampling or permutation (see section 1.4). Note that we scale $\widehat{\beta}_j^{(m)}$ by $\widetilde{\sigma}_j^{(m)}$, which is an estimated standard error of $n^{\frac{1}{2}}(\widetilde{\beta}_j^{(m)} - \beta_{0j}^{(m)})$, since under $H_j^{(m)}$, $\widehat{\sigma}_j^{(m)} = \text{Var}\{n^{\frac{1}{2}}(\widehat{\beta}_j^{(m)} - \beta_{0j}^{(m)})\}^{1/2} \to 0$ and the null distribution of $n^{\frac{1}{2}}\widehat{\beta}_j^{(m)}/\widehat{\sigma}_j^{(m)}$ is difficult to approximate.

To test $\mathcal{H}_j$ simultaneously, we order the test statistics $\mathbf{t}_j = (t_j^{(1)}, ..., t_j^{(M)})^\intercal$ from largest to smallest — $t_j^{(r_1)} \geq t_j^{(r_2)} \geq ... \geq t_j^{(r_M)}$ — and identify their corresponding hypotheses $H_j^{(r_1)}, ..., H_j^{(r_M)}$. Define for every $\Omega \subset \{1, ..., M\}$ the sup-statistic over Ω and its corresponding reference distribution: $s_j^\Omega = \max_{m\in\Omega} t_j^{(m)}$ and $\mathcal{S}_j^\Omega = \{\max_{m\in\Omega} t_j^{*b(m)}\}_{b=1,..,B}$. Furthermore, denote the ψth quantile of $\mathcal{S}_j^\Omega$ by $c_j^\Omega(\psi)$, which approximates the ψth quantile of s_j^Ω under

the null that $\{\beta_j^{(m)} = 0 : m \in \Omega\}$. We identify the subset of hypotheses to reject, denoted by $\mathscr{R}_j$, as follows.

1) Let $\Omega_1 = \{1, ..., M\}$. If $s_j^{\Omega_1} \leq c_j^{\Omega_1}(\psi)$, accept all hypotheses and stop. Otherwise, let $\mathscr{R}_j = \{r_1\}$ and continue.

   $\vdots$

k) Let $\Omega_k = \Omega_1 \setminus \mathscr{R}_j$. If $s_j^{\Omega_k} \leq c_j^{\Omega_k}(\psi)$, accept all hypotheses in $\{H_j^{(m)}\}_{m \in \Omega_k}$ and stop. Otherwise, let $\mathscr{R}_j = \mathscr{R}_j \cup \{r_k\}$ and continue.

   $\vdots$

M) Let $\Omega_M = \{r_M\}$. If $s_j^{\Omega_M} \leq c_j^{r_M}(\psi)$, accept $H_j^{(r_M)}$. Otherwise, let $\mathscr{R}_j = \mathscr{R}_j \cup \{r_M\}$.

The step down procedure for the simultaneous testing of $\mathcal{H}_j$ then rejects all hypotheses in $\{H_j^{(m)}\}_{m \in \mathscr{R}_j}$ and concludes that x_j is associated with $\{y^{(m)}\}_{m \in \mathscr{R}_j}$. If the probability of making a type I error at each step is α

$$P\left(s_j^{\Omega_k} > c_j^{\Omega_k}(\psi) \Big| \bigcap_{m \in \Omega_k} H_j^{(m)}\right) = \alpha, \qquad \text{for any } k$$

then the FWER of the stepdown procedure – that is, the probability of making at least one false rejection over the set $\mathcal{H}_j$ – is maintained at α.

One of the main results of this paper is that, given a suitably estimated $\widehat{\boldsymbol{\beta}}$, the FWER of our stepdown procedure approaches 0 as $n \to \infty$ regardless of what quantile ψ we use to determine the cutoff for rejection. Specifically, define $\mathcal{A}$ and $\mathcal{A}^c$ as indexing the non-zero and zero components of $\boldsymbol{\beta}_0$, respectively, where $\boldsymbol{\beta}_{\mathcal{A}}$ denotes the subvector of $\boldsymbol{\beta}$ corresponding to $\mathcal{A}$. Furthermore, define a *sparsistent* estimator $\widehat{\boldsymbol{\beta}}$ to be one that satisfies $P(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$ as $n \to \infty$. Then we have the following result.

Theorem 1. *If $\widehat{\boldsymbol{\beta}}$ is sparsistent, then $P\left(s_j^{\Omega} > c_j^{\Omega}(\psi) \big| \bigcap_{m \in \Omega} H_j^{(m)}\right) \to 0$ as $n \to \infty$ for any Ω and ψ, and SMRT has an asymptotic familywise error rate of 0.*

The proof is given in Appendix 1.7.1. The result follows from the fact that, under a given null $\bigcap_{m \in \Omega} H_j^{(m)}$, the test statistic s_j^{Ω} is estimated at exactly 0 with probability tending to 1. If $s_j^{\Omega} = 0$, we cannot reject $\bigcap_{m \in \Omega} H_j^{(m)}$, regardless of the value of $c_j^{\Omega}(\psi)$. Of course,

in finite samples, choice of ψ is paramount in maintaining a desired error rate. We discuss this choice and other tuning issues in section 1.4. In section 1.3, we discuss our estimation procedure which ensures sparsistency.

1.2.2 Controlling error rates across all predictors

To extend FWER control to the set of all hypotheses $\{H_j^{(m)}\}_{j=1,...,p;m=1,..,M}$, one could test each $\mathcal{H}_j$ at level α/p. This may be a conservative strategy, as it relies on a union-bound argument. Instead, we simply employ a stepdown procedure on all of the test statistics $\mathbf{t} = (\mathbf{t}_1^\intercal, ..., \mathbf{t}_p^\intercal)^\intercal$. Let $\mathbf{u} = (u_1, ..., u_T)^\intercal = \mathbf{t}$ and $\widetilde{\mathcal{H}} = \{\widetilde{H}_1, ..., \widetilde{H}_T\} = \{H_1^{(1)}, ..., H_1^{(M)}, H_2^{(1)}, ..., H_p^{(M)}\}$ be relabeled test statistics and hypotheses, respectively, where $T = Mp$. Furthermore, define $s_\Omega = \max_{i \in T} u_i$ and define S_Ω and $c_\Omega(\psi)$ analogously to in the previous section. And finally let $\mathscr{R}$ be the set of rejected hypotheses. Then the stepdown procedure proceeds exactly as for a single predictor and in the end we reject all hypotheses $\{\widetilde{H}_i\}_{i \in \mathscr{R}}$.

1.3 Estimation

The aforementioned testing procedure relies on a sparse regularized estimator $\widehat{\boldsymbol{\beta}}$ and its asymptotic properties. We next detail the construction of $\widehat{\boldsymbol{\beta}}$ as well as the asymptotic distribution for the zero and non-zero components, which is crucial for the justification of our proposed testing procedures. Specifically, we simultaneously consider all M outcomes and obtain $\widehat{\boldsymbol{\beta}}$ as the minimizer of the penalized sum of negative PLLs

$$-\sum_{m=1}^{M} \mathcal{L}^{(m)}(\boldsymbol{\beta}^{(m)}) + p_{\lambda,\mathbf{w}}(\boldsymbol{\beta}) \tag{1.3}$$

where the penalty function

$$p_{\lambda,\mathbf{w}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} d_j + \lambda \sum_{m=1}^{M} \sum_{j=1}^{p} w_j^{(m)} \left|\alpha_j^{(m)}\right|, \text{ with } \beta_j^{(m)} = d_j \alpha_j^{(m)}, \text{ subject to } d_j \geq 0,$$

corresponds to the adaptive hierarchical lasso penalty proposed in Zhou and Zhu (2010) for grouped predictor variables, penalty parameter λ which controls the amount of regularization, and weight $w_j^{(m)} = |\widetilde{\beta}_j^{(m)}|^{-1}$ chosen to ensure oracle properties of $\widehat{\boldsymbol{\beta}}$. Summing over

the PLLs in (1.3) essentially imposes a working independence assumption across the outcomes (Liang and Zeger, 1986). Incorporating covariance information about $\mathbf{y}$ can improve efficiency, which we discuss further in section 1.6. On the other hand, imposing the joint penalty $p_{\lambda,\mathbf{w}}(\boldsymbol{\beta})$ incorporates the potential for joint sparsity across all outcomes for some x_j's. Setting $d_j = 0$ declares x_j to be non-informative for all outcomes or equivalently $\boldsymbol{\beta}_{0j}$ is 0; while setting $\alpha_j^{(m)} = 0$ suggests that $\beta_{0j}^{(m)} = 0$.

Now, since $\{\mathcal{L}^{(m)}\}_{m=1,...,M}$ are non-linear functions without closed form in most cases, direct maximization of (1.3) may be numerically challenging, especially when p is not small. To overcome these difficulties in practice, we propose to take a quadratic expansion of $\mathcal{L}^{(m)}(\boldsymbol{\beta})$ in (1.3) similar to Zhang and Lu (2007); Wang and Leng (2007). Specifically, we instead minimize $\sum_{m=1}^{M}(\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal}\widetilde{\mathbf{I}}^{(m)}(\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)}) + p_{\lambda,\mathbf{w}}(\boldsymbol{\beta})$, which can be re-written as

$$\|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2 + p_{\lambda,\mathbf{w}}(\boldsymbol{\beta}), \tag{1.4}$$

where $\widetilde{\mathbf{I}}^{(m)} = -\ddot{\mathcal{L}}^{(m)}(\widetilde{\boldsymbol{\beta}}^{(m)})$, $\ddot{\mathcal{L}}^{(m)}(\mathbf{b}) = \partial^2\mathcal{L}^{(m)}(\mathbf{b})/\partial\mathbf{b}\partial\mathbf{b}^{\intercal}$, $\widetilde{\mathbb{X}} = \mathrm{diag}(\widetilde{\boldsymbol{\Lambda}}^{(1)}, ..., \widetilde{\boldsymbol{\Lambda}}^{(M)})$, $\widetilde{\mathbf{Y}} = \widetilde{\mathbb{X}}\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\Lambda}}^{(m)}$ is a symmetric half matrix of $\widetilde{\mathbf{I}}^{(m)}$ such that $\widetilde{\mathbf{I}}^{(m)} = \widetilde{\boldsymbol{\Lambda}}^{(m)}\widetilde{\boldsymbol{\Lambda}}^{(m)}$. Thus, we have reduced the original complicated minimization problem into a penalized L_2 minimization problem, which is much more tractable using widely available software. Computational simplifications and a full algorithm for fitting are discussed in appendix 1.7.4.

1.3.1 Asymptotic Theory

In this section, we present the properties of our proposed estimator $\widehat{\boldsymbol{\beta}}$. It is sparsistent in that it asymptotically sets truly null effects to exactly 0, and our estimates of non-null effects are asymptotically normal and possess the *oracle property*, in that they are as efficient in the limit as if we knew which effects were truly null *a priori*. Let $\mathbf{I}_{\mathcal{A},\mathcal{B}}$ denotes the submatrix of $\mathbf{I}$ corresponding to rows in $\mathcal{A}$ and columns in $\mathcal{B}$.

In Appendix 1.7.2, we show that for PLLs $\{\mathcal{L}^{(m)}(\boldsymbol{\beta}^{(m)})\}_{m=1,...,M}$ that satisfy the regularity conditions (also listed in the appendix), if $n^{-1}\sqrt{\lambda} = o_p(n^{-1/2})$, then there exists a root-n consistent local maximizer $\widehat{\boldsymbol{\beta}}$ such that $P(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$ and

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \to N(0, \mathbf{I}_{\mathcal{A},\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}}\mathbf{I}_{\mathcal{A},\mathcal{A}}^{-1})$$

in distribution, where $\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{A}} = \text{Cov}(\boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0))$, $\boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}})$ denotes the contribution of the ith subject to the profile score function for $\boldsymbol{\beta}_{\mathcal{A}}$, $\mathbf{I} = \text{diag}\{\mathbf{I}^{(1)}, ..., \mathbf{I}^{(m)}\}$, and $\mathbf{I}^{(m)}$ is the negative limiting information matrix corresponding to the profile likelihood for $\boldsymbol{\beta}_0^{(m)}$.

This result, parallel to that given in Zhou and Zhu (2010), offers the promise of identifying null effects with probability approaching 1, while still efficiently estimating non-null effects. From a testing perspective, it motivates theorem 1, and ensures the probability of making a type I error in SMRT decreases to 0 as $n \to \infty$.

1.3.2 Estimating the variability in $\widehat{\boldsymbol{\beta}}$

The asymptotic results on $\widehat{\boldsymbol{\beta}}$ suggest that we are as efficient in the limit as if we knew which parameters were truly 0 from the outset. However, in finite samples the added variability due to estimating $\mathcal{A}^c$ may not be negligible, and hence relying on the asymptotic result will underestimate the variability in $\widehat{\boldsymbol{\beta}}$. To better approximate the finite sample distribution, we propose a perturbation resampling procedure to estimate the distribution of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. This procedure, by accounting for the variability in estimating $\mathcal{A}^c$, provides a more precise estimate of the variability in $\widehat{\boldsymbol{\beta}}$, and maintains the correlation structure in $\widehat{\boldsymbol{\beta}}$.

We generate a resampled counterpart of $\widehat{\boldsymbol{\beta}}$, denoted by $\widehat{\boldsymbol{\beta}}^*$, in two steps. We first generate $\widetilde{\boldsymbol{\beta}}^*$, a resampled version of $\widetilde{\boldsymbol{\beta}}$, by either perturbing the non-parametric likelihood or directly perturbing the influence function corresponding to $\widetilde{\boldsymbol{\beta}}$. Then we minimize our objective function (1.4) using $\widetilde{\boldsymbol{\beta}}^*$ in place of $\widetilde{\boldsymbol{\beta}}$, yielding resampled estimates $\widehat{\boldsymbol{\beta}}^*$. Specifically, let $\mathcal{G} = (G_1, ..., G_n)^\intercal$ be a vector of iid positive random variables with $E[G_i] = 1$ and $\text{Var}(G_i) = 1$, generated independently of the data. We obtain $\widetilde{\boldsymbol{\beta}}^*$ as the minimizer of $\sum_{m=1}^{M} \mathcal{L}^{(m)*}(\boldsymbol{\beta}^{(m)})$ or explicitly as

$$\widetilde{\boldsymbol{\beta}}^* = \widetilde{\boldsymbol{\beta}} + \sum_{i=1}^{n} \widetilde{\mathbf{I}}^{-1} \widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})(G_i - 1)$$

where $\mathcal{L}^{(m)*}(\boldsymbol{\beta})$ is the profile likelihood corresponding to the perturbed non-parametric likelihood with the contribution of the ith subject weighted by G_i, $\widetilde{\mathbf{I}}$ is the observed information matrix for $\boldsymbol{\beta}$ evaluated at $\widetilde{\boldsymbol{\beta}}$, and $\widetilde{\boldsymbol{\varphi}}_i(\boldsymbol{\beta})$ is the empirical estimate of the score function $\boldsymbol{\varphi}_i(\boldsymbol{\beta})$. In the second step, we find $\widehat{\boldsymbol{\beta}}^*$ as the minimizer of $Q^*(\boldsymbol{\beta}) = \sum_{m=1}^{M} (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{*(m)})^\intercal \widetilde{\mathbf{I}}^{(m)} (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{*(m)}) + \sum_{j=1}^{p} d_j + \lambda \sum_{m=1}^{M} \sum_{j=1}^{p} w_j^{*(m)} |\alpha_j^{(m)}|$ subject to $d_j \geq 0$, $w_j^{*(m)} = |\widetilde{\beta}_j^{*(m)}|^{-1}$. Similar

resampling procedures have been proposed for making inference with a wide range of standard objective functions without regularization (Jin et al., 2001; Tian et al., 2007; Uno et al., 2007, e.g) and recently extended to accommodate L_1-type regularized estimators (Minnier et al., 2011). Here, we propose such a resampling procedure to both account for the potential correlation among the outcomes and better approximate the finite sample behavior of hierarchically regularized estimators. Our second main result concerns the asymptotic properties of the resampled $\widehat{\boldsymbol{\beta}}^*$.

Theorem 2. *For PLLs* $\{\mathcal{L}^{(m)}(\boldsymbol{\beta}^{(m)})\}_{m=1,\ldots,M}$ *that satisfy the regularity conditions listed in the appendix, if* $n^{-1}\sqrt{\lambda} = o_p(n^{-1/2})$*, then there exists a local maximizer* $\widehat{\boldsymbol{\beta}}^*$ *of* $Q^*(\boldsymbol{\beta})$ *such that*

(i) $\|\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

(ii) $P\left(\widehat{\boldsymbol{\beta}}^*_{\mathcal{A}^c} = 0 | \mathbb{V}\right) \to 1$ *as* $n \to \infty$

(iii) $n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\beta}}^*_{\mathcal{A}} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}}\right) \mid \mathbb{V}$ *converges in distribution to* $N(0, \mathbf{I}^{-1}_{\mathcal{A}\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{I}^{-1}_{\mathcal{A}\mathcal{A}})$

The proof is given in appendix 1.7.3. This theorem indicates that, given the observed data, $n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\beta}}^*_{\mathcal{A}} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}}\right)$ has the same limiting distribution as $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}})$. Thus, to approximate the distribution of $\widehat{\boldsymbol{\beta}}$ for a given dataset, we may generate a large number, say B, of $\mathcal{G}$, denoted by $\{\mathcal{G}^{[b]}\}_{b=1,\ldots,B}$. Let $\widehat{\boldsymbol{\beta}}^{*_b}$ denote the corresponding realization of $\widehat{\boldsymbol{\beta}}^*$. To construct a confidence interval (CI) for a specific $\beta_j^{(m)}$, one may estimate the standard error of $\widehat{\beta}_j^{(m)}$ as $\widehat{\sigma}_j^{(m)}$ the empirical standard error of its perturbed realizations, $\left\{\widehat{\beta}_j^{*_b(m)}\right\}_{b=1,\ldots,B}$. An $100(1-\alpha)\%$ level confidence interval can then be constructed based on the normal confidence interval $\widehat{\beta}_j^{(m)} \pm \mathcal{Z}_{1-\alpha/2}\widehat{\sigma}_j^{(m)}$ or alternatively the lower and upper $\alpha/2$ percentiles of $\left\{\widehat{\beta}_j^{*_b(m)}\right\}_{b=1,\ldots,B}$.

1.4 Tuning

In this section, we discuss issues relating to tuning for estimation and testing. First, we discuss choosing the reference distribution for $t_j^{(m)}$ for use in testing. Next, we consider the choice of ψ, which determines the cut-off for rejection in the stepdown procedure. Finally, we discuss choosing the penalization tuning parameter λ.

1.4.1 Choosing a reference distribution

The reference distribution $\mathcal{T}_j^{(m)} = \{t_j^{*_b(m)}\}_{b=1,\ldots,B}$ needs to approximate the distribution of $t_j^{(m)} = n^{\frac{1}{2}} \left|\widehat{\beta}_j^{(m)}\right| / \widetilde{\sigma}_j^{(m)}$ under $H_j^{(m)}$. An immediately appealing choice for the reference distribution is to use the resampled $\widehat{\boldsymbol{\beta}}^*$, since, as we stated in section 1.3.2, $n^{\frac{1}{2}} \left(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^* - \widehat{\boldsymbol{\beta}}_{\mathcal{A}}\right) \mid \mathbb{V} \longrightarrow N(0, \mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1} \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} \mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1})$. Thus, we could choose $t_j^{*_b(m)} = n^{\frac{1}{2}} \left|\widehat{\beta}_j^{*_b(m)} - \widehat{\beta}_j^{(m)}\right| / \widetilde{\sigma}_j^{(m)}$. However, although resampling provides good approximation to the finite sample distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$, it tends to over-estimate the variability of $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}$. As a result, testing based on $\mathcal{T}_j^{(m)}$ may be conservative if we choose $\psi = 1 - \alpha$. One approach to overcome this is to set ψ at a lower (less conservative) level – although empirically identifying a proper ψ to both preserve the type I error and achieve high power is challenging.

Alternatively, we propose the use of permutation to better approximate the null distributions. Let $\widehat{\boldsymbol{\beta}}(\Omega) = (\widehat{\beta}_j^{(m)}(\Omega))_{j=1,\ldots,p;m=1,\ldots,M}$ denote the estimate of $\boldsymbol{\beta}_0$ using the dataset $\{(\mathbf{y}_i^{\Omega^\intercal}, \mathbf{x}_i^\intercal)^\intercal\}_{i=1,\ldots,n}$, where $\{\mathbf{y}_i^\Omega\}_{i=1,\ldots,n}$ denotes a partially permuted counterpart of $\{\mathbf{y}_i\}_{i=1,\ldots n}$ with $\{y_i^{(m)}\}_{m\in\Omega;i=1,\ldots,n}$ randomly permuted across subjects but $\{y_i^{(m)}\}_{m\notin\Omega;i=1,\ldots,n}$ unchanged. And let $\widehat{\boldsymbol{\beta}}^b(\Omega)$ be the bth such permutation-based estimate. To be clear, for example, $\widehat{\boldsymbol{\beta}}^b(\{1\})$ corresponds to the estimate of $\boldsymbol{\beta}_0$ from a dataset where only the first outcome $\{y_i^{(1)}\}_{i=1,\ldots,n}$ is permuted.

The reference distribution that we pursue in our simulations is a composite distribution obtained by permuting each of the outcomes individually. For each m, we obtain $\left\{\widehat{\boldsymbol{\beta}}^b(\{m\})\right\}_{b=1,..,B}$ and retain only those elements which pertain to outcome m: $\left\{\widehat{\beta}_j^{b(m)}(\{m\})\right\}_{j=1,\ldots,p;b=1,\ldots,B}$. We then define the reference distribution for the stepdown procedure as

$$t_j^{*_b(m)} = n^{\frac{1}{2}} \left|\widehat{\beta}_j^{b(m)}(\{m\})\right| / \widetilde{\sigma}_j^{(m)}.$$

In this way, we are essentially obtaining a reference distribution for (a standardized) $\widehat{\boldsymbol{\beta}}_j^{(m)}$ under the null hypothesis $\bigcap_{j=1,\ldots,p} H_j^{(m)}$. This strategy has the undesirable consequence of breaking the correlation structure in $\mathbf{y}$, since $t_j^{*_b(m)}$ and $t_j^{*_b(m')}$ are obtained under different permutation regimes for $m \neq m'$. But defining $t_j^{*_b(m)}$ in this way allows us to approximate the distribution of $\widehat{\beta}_j^{(m)}$ without making any assumption about $\bigcap_{m'\neq m} H_j^{(m')}$. This permutation distribution, while not guaranteeing exact control of the FWER, does provide asymptotic

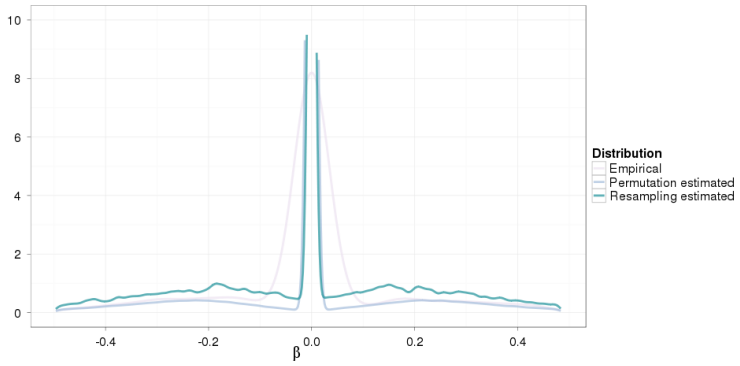


Figure 1.2: Simulation-based empirical and estimated distribution of null effects. Empirical null distribution of $\widehat{\beta}_j^{(m)}$ (light purple) agrees closely in the tails with the permutation-based estimate (dark purple), while the resampling-based estimate (blue-green) overestimates the density in the tails.

control of the FWER. In our numerical studies, we observe that it does a much better job of approximating the null distribution of the regularized estimator $\widehat{\boldsymbol{\beta}}$ compared to that obtained based on resampling, as evidenced by figure 1.2.

1.4.2 Choosing ψ

In order to control the FWER at the level α for predictor x_j with associated hypotheses $\mathcal{H}_j$, one can choose $\psi = 1 - \alpha$. This ensures that $P\left(s_j^{\Omega_k} > c_j^{\Omega_k}(\psi) \middle| \bigcap_{m \in \Omega_k} H_j^{(m)}\right) = \alpha$, for any k, which in turn ensures FWER control. Similarly, in order to control the FWER across all p predictors using the union-bound, one can choose $\psi = 1 - \alpha/p$ for each $\mathcal{H}_j$ or perform the full stepdown procedure on $\mathbf{u}$ with $\psi = 1 - \alpha$.

If one wanted to use the resampling-based reference distribution with

$$t_j^{*b(m)} = n^{\frac{1}{2}} \left| \widehat{\beta}_j^{*b(m)} - \widehat{\beta}_j^{(m)} \right| / \widetilde{\sigma}_j^{(m)},$$

it is possible to correct for its conservativeness by choosing $\psi < 1 - \alpha$, while still maintaining the level α. One could use permutation methods or another layer of resampling to estimate the smallest ψ that would still maintain the level α. However, that requires computing a large number of resamples within a large number of permutations or resamples, which becomes prohibitively computationally demanding quickly.

1.4.3 Choosing λ

SMRT involves a large number of model fits — an equation like (1.4) is minimized once for each resample or permutation. Each one of those minimizations requires the choice of the lasso tuning parameter λ. Because of the large number of fits, using a time-consuming method like cross-validation to choose λ is not feasible. We propose to use a BIC-style criteria:

$$\lambda = \underset{\lambda}{\operatorname{argmin}} \left(\|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}\boldsymbol{\beta}_\lambda\|_2^2 + \min\{n^{0.1}, \log n\}\mathrm{df}_\lambda \right),$$

where $\boldsymbol{\beta}_\lambda$ is the minimizer of (1.4) corresponding to λ and df_λ is the number of non-zero entries in $\boldsymbol{\beta}_\lambda$. In small and moderate sample sizes, $n^{0.1}$ is much smaller than $\log n$ and is used here. However, when n becomes large $\log n$ may be preferred.

1.5 Example

1.5.1 Genetic study to identify shared autoimmune risk loci

We apply our SMRT to a study of shared autoimmunity with the goal of identifying genetic markers associated with 4 autoantibodies: anti-nuclear antibodies (ANA), anti-cyclic citrullinated protein (CCP) antibodies, anti-transglutaminase (TTG) antibodies, and anti-thyroid peroxidase antibodies (TPO). These 4 autoantibodies are respectively markers for 4 autoimmune diseases (ADs): systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), celiac disease and autoimmune thyroid disease. The genetic markers consists of 67 single-nucleotide polymorphisms (SNPs) previously published as potential risk markers for these four ADs. Discovering which SNPs regulate multiple ADs can aid in understanding potential shared pathways or etiology of these diseases (Zhernakova et al., 2009). While the co-occurrence of multiple ADs within individuals has been documented (Somers et al., 2006), it would be rare, even for someone who is at high risk for the spectrum of ADs, to have more than one. In contrast, autoantibodies can be present in individuals predisposed to having the disease even in the absence of a disease phenotype. For example, while co-occurrence of ADs within families is well documented (Somers et al., 2006), family members of those with autoimmune disease may also experience elevated levels of autoantibodies if they haven't (yet)

exhibited the disease phenotype. In this study, the autoantibodies are considered markers for subjects at higher risk for SLE, celiac and autoimmune thyroid disease.

The study cohort includes 1265 individuals of European ancestry with RA identified through electronic medical records at Partners Healthcare (Liao et al., 2010; Kurreeman et al., 2011). Due to a limit of quantification, the antibody measurements are highly unreliable when the values are either very low or very high. A convenient approach to incorporating such limitations is by assuming a marginal proportional odds model and truncating the observations at the limit of quantification, which corresponds to (1.1) with $\epsilon^{(m)}$ coming from a logistic distribution. Hence $\beta_{0j}^{(m)}$ still has the interpretation of being a log odds ratio (OR).

Results for the autoantibody data are summarized in figure 1.3. Figure 1.3 (a) shows results for the sparse estimation step. In the figure, SNPs are denoted along the y-axis, and outcomes are denoted along the x-axis. Color of the tile indicates the OR estimate. The color scale indicates strength and direction of estimated association, with darker red (blue) colors indicating more positive (negative) association and white indicating no association. In order to measure the strength of association with respect to the FWER, we adapt the notion of a q-value (Storey, 2003) for each test to be the smallest α such that that test would reject while controlling the FWER for that SNP at α. Figure 1.3 (b) shows this q-value for each test.

For such a large number of total hypotheses, we do not have sufficient power to detect multiple regulation while simultaneously controlling the familywise error rate across all SNPs. Taking a less conservative view, if we control the FWER at the SNP level, five SNPs show some evidence of multiple regulation at $\alpha = 0.1$. The SNP rs2187668, which had previously shown associations to SLE (Taylor et al., 2011) and celiac (van Heel et al., 2007), we estimate to be related to the autoantibodies for those diseases at OR = 1.45 (q-value 0.005) for ANA and OR = 1.62 (q-value 0.005) for TTG, as well as to CCP (OR = 0.78, q-value 0.05). This SNP is in the MHC region, which is known to affect immune function. Similarly, rs3129860, also in the MHC region, which had previously shown an association to SLE (Taylor et al., 2011), here demonstrated an association to ANA (OR = 1.28, q-value 0.05), CCP (OR = 1.50, q-value 0.003), and TPO (OR = 1.30, q-value 0.05). After having previously been

associated with both autoimmune thyroid disease (Ueda et al., 2003) and RA (Plenge et al., 2005), rs3087243, which is located in the CTLA4 gene, was found to be associated with CCP (OR = 1.18, q-value 0.06) and TTG (OR = 1.22, q-value 0.07). And rs4963128 was similarly associated with CCP (OR = 0.85, q-value 0.06) and TTG (OR = 1.23, q-value 0.06), having been associated with SLE in previous studies (Harley et al., 2008). Finally, rs6679677, located on gene RSBN1, displayed as association with CCP (OR = 1.32, q-value = 0.02) and TTG (OR = 1.29, q-value = 0.10), after having previously been identified as associated with RA (Burton et al., 2007) and hyperthyroidism (Eriksson et al., 2012).

1.5.2 Simulation results

We ran simulations to assess the performance of our point and interval estimation procedures as well as SMRT. We loosely based our simulations on the autoantibody dataset, allowing the relationship between $\mathbf{x}$ and $\mathbf{y}$ to be specified by a proportional odds model. We considered sample sizes of 150, 250, and 500 and ran 1000 simulations for each sample size. For each simulation, 1000 resampled $\widehat{\boldsymbol{\beta}}^*$s were generated.

We set the number of predictors of interest p to be 30 and the number of outcomes M to be 4. Covariates $\mathbf{x}$ took values in $\{0, 1, 2\}$ with probability $\{p^2, 2p(1-p), (1-p)^2\}$ where $p = 0.15$. Outcomes $\mathbf{y}$ were generated according to the marginal proportional odds model, conditional on $\mathbf{x}$. We allowed correlation in $\mathbf{y}$, which was accomplished by first generating correlated normal random variables $\mathbf{z}_i \sim N_4(\mathbf{0}, \Sigma)$ where $\Sigma = 0.85I + 0.15\mathbf{1}\mathbf{1}^\intercal$ is exchangeable. Then let $\mathbf{u}_i = \Phi(\mathbf{z}_i)$ for Gaussian distribution function $\Phi(\cdot)$, and finally

$$\mathbf{y}_i = \exp\{\mathbf{x}_i\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_i\}$$

where $\boldsymbol{\epsilon}_i = \log(\frac{u_i}{1-u_i}) \sim$ logistic. For computational simplicity, we discretized $\mathbf{y}$ into ten levels.

The relationship between $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$[\boldsymbol{\beta}_0^{(1)}, ..., \boldsymbol{\beta}_0^{(M)}] = \begin{bmatrix} \mathbf{1}_{20} & \frac{\mathbf{1}}{\mathbf{2}}_{16} & \mathbf{1}_{12} & \frac{\mathbf{1}}{\mathbf{2}}_{8} \\ \mathbf{0}_{10} & \mathbf{0}_{14} & \mathbf{0}_{18} & \mathbf{0}_{22} \end{bmatrix}_{30\times 4}.$$

where $\mathbf{1}_k$ is a $k \times 1$ vector of ones, $\mathbf{0}_k = 0 \times \mathbf{1}_k$ and $\frac{\mathbf{1}}{\mathbf{2}}_k = \frac{1}{2} \times \mathbf{1}_k$.

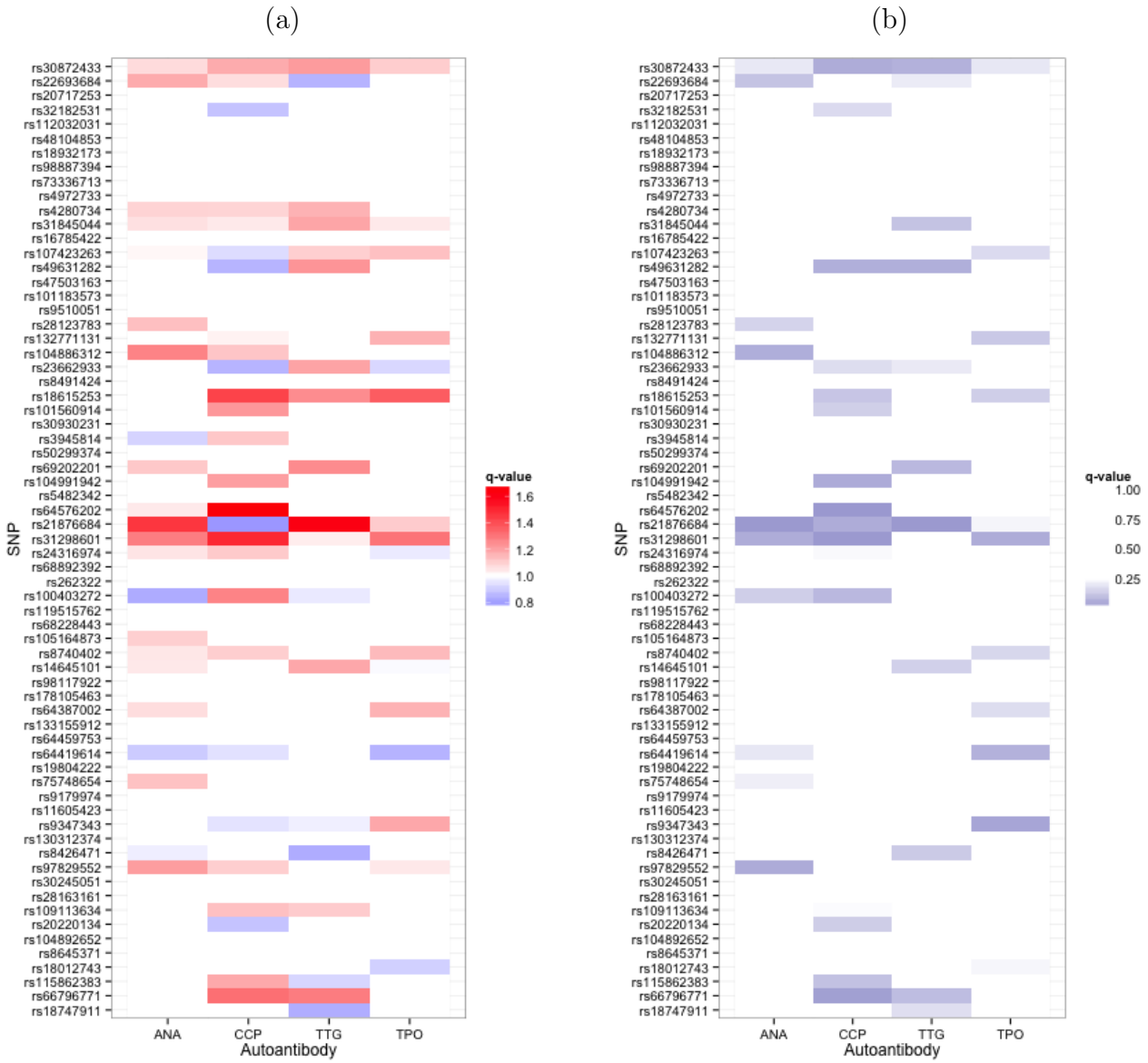


Figure 1.3: Results for autoantibody data. SNPs are listed on the y-axis, and autoantibodies are listed on the x-axis. (a) Sparse effect estimates. The color of the rectangle indicates the effect. Red indicates positive association (OR > 1) between SNP and autoantibody; blue indicates negative association (OR < 1). Darker colors indicate larger magnitudes, and white indicates no estimated association. (b) q-values. Darker color indicates smaller q-value and more evidence against the null hypothesis of no association.

This configuration indicates that there are eight predictors related to all four outcomes, four related to just the first three outcomes, four related to just the first two outcomes, and four related to just the first outcome. The remaining ten predictors are null, unrelated to any outcome. We also see that associations to outcomes $y^{(2)}$ and $y^{(4)}$ are weak, so we would expect there to be less power to detect those effects.

1.5.3 Bias, standard errors (SEs), and confidence intervals (CIs)

We first demonstrate that our point and interval estimation procedures perform well in finite samples. Figure 1.4 shows the average bias in $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ across simulations, plotted according to true effect size $\boldsymbol{\beta}_0$ and sample size. The regularized $\widehat{\boldsymbol{\beta}}$ exhibits much smaller bias than the unregularized $\widetilde{\boldsymbol{\beta}}$ for all sample sizes and effect sizes. Particularly at smaller sample sizes, regularization substantially reduces the bias in the estimator.

In figure 1.5, we plot the average percent bias in SE estimates obtained based on our proposed resampling procedures as well as those based on the asymptotic variance. Both the asymptotic SE estimate and the resampling-based one $\widehat{\sigma}_j^{(m)}$ overestimate the variability in $\widehat{\beta}_j^{(m)}$ when $\beta_{0j}^{(m)} = 0$, but $\widehat{\sigma}_j^{(m)}$ more closely approximates $\sigma_j^{(m)}$. When $\beta_{0j}^{(m)} \neq 0$, the asymptotic SE tends to underestimate the true variability, while $\widehat{\sigma}_j^{(m)}$ approximates it well.

We examine CI coverage in figure 1.6 and see that underestimating the SEs leads to poor 95% CI coverage levels for the normal-based CI methods, based on $\widetilde{\sigma}_j^{(m)}$ and $\widehat{\sigma}_j^{(m)}$. Resampling-based quantile 95% CIs have good coverage for all values of $\beta_{0j}^{(m)}$ and all sample sizes. The coverage levels of asymptotic-based CIs are as low as 78% for non-zero effects and remain lower than the nominal level even when $n = 500$. Hence in practice, we recommend the quantile based CIs.

1.5.4 Testing

In this section, we examine the performance of our proposed testing procedures. To demonstrate the role of regularization in improving testing, we compare SMRT to an unregularized version, which we will call MRT. Throughout we use the permutation method outlined in section 1.4.1 as the reference distribution for both SMRT and MRT and take $\psi = 1 - \alpha$. To

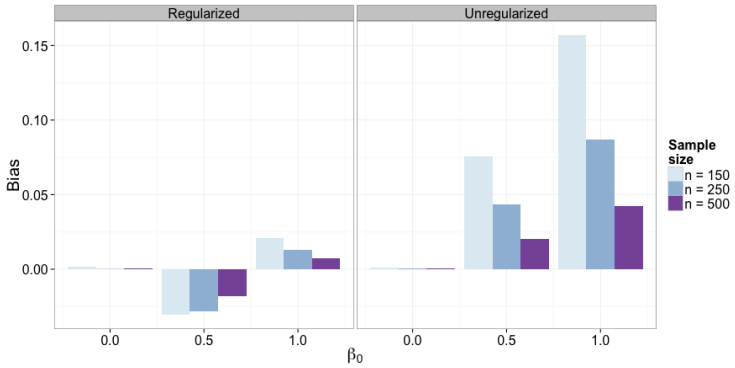


Figure 1.4: Average estimated bias in regularized $\widehat{\beta}_j^{(m)}$ and unregularized $\widetilde{\beta}_j^{(m)}$ across 1000 simulations plotted against $\beta_{0j}^{(m)}$. Results for $\widehat{\boldsymbol{\beta}}$ are depicted on the left and for $\widetilde{\boldsymbol{\beta}}$ on the right. Color denotes sample size. Bias estimates are aggregated over all estimates that share the same effect size.

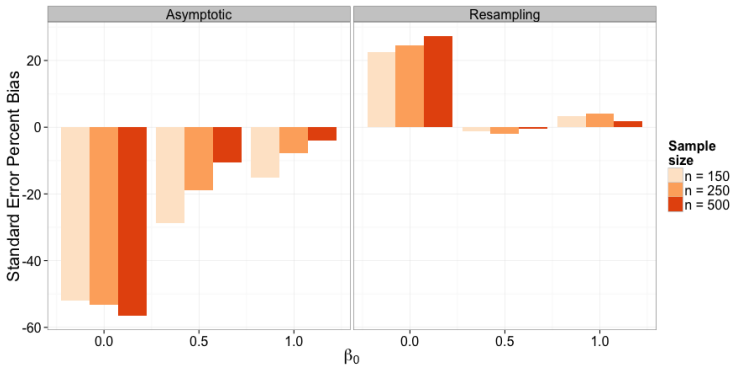


Figure 1.5: Average estimated bias of estimates of $\sigma_j^{(m)}$ the SE of $\widehat{\beta}_j^{(m)}$ across 1000 simulations, arranged according to value of $\beta_{0j}^{(m)}$. Results for the asymptotic SE are on the left, and those for the resampling-based $\widehat{\sigma}_j^{(m)}$ are on the right. Bias was estimated as the difference between the empirical SE of $\widehat{\beta}_j^{(m)}$ across simulations and the estimated one. Bias estimates are aggregated over all estimates that share the same effect size. Color denotes sample size.

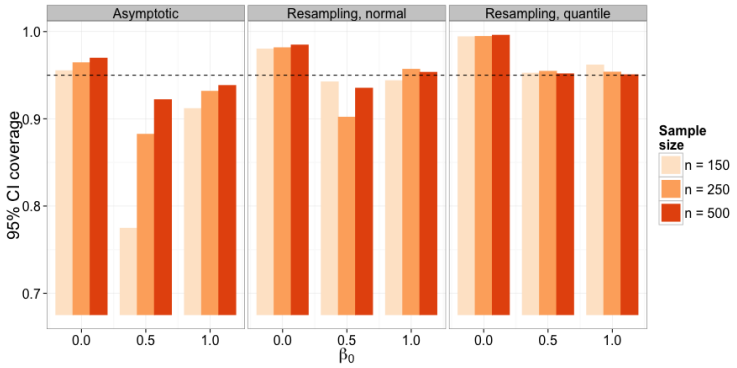


Figure 1.6: 95% CI coverage across 1000 simulations arranged according to value of $\beta_{0j}^{(m)}$. Results for normal-based CIs based on asymptotic SEs are depicted in the left panel, normal-based CIs based on $\widehat{\sigma}_j^{(m)}$ in the middle panel, and CIs based on quantiles of $\widehat{\beta}_j^{*(m)}$ in the right panel. Coverage estimates are aggregated over all estimates that share the same effect size. Color denotes sample size.

demonstrate the advantages of using the stepdown method, we compare also to a single-step procedure, which we will denote Sup in the following, where we reject all $H_j^{(m)}$ for which $t_j^{(m)} > c_j^{\Omega_1}(\psi)$, where as before $\Omega_1 = \{1, ..., M\}$. Finally, we compare to the Bonferroni adjustment by computing the asymptotic p-values $p_j^{(m)} = 2\Phi\left(-\left|\widehat{z}_j^{(m)}\right|\right)$ and rejecting $H_j^{(m)}$ if $p_j^{(m)} < \frac{\alpha}{Mp}$.

When controlling the FWER at $\alpha = 0.05$ for each SNP, SMRT and MRT performed similarly in controlling FWER. For $n = 250$, the average empirical FWER was 0.055 for SMRT and MRT. The more conservative Sup test had average FWER of 0.039, and the even more conservative Bonferroni 0.003. All three of SMRT, MRT, and the Sup test had empirical FWER for one SNP as high as 0.07 (0.077 for MRT, 0.072 for SMRT and Sup). Results for other sample sizes were similar.

In terms of power, SMRT dominates all other test procedures. Figure 1.7 depicts the power to detect non-null effects at $n = 250$ (other sample sizes show similar relative performances, with SMRT performing relatively better as sample size decreases). Tests of the form $\{H_0 : \bigcup \beta_j^{(m)} = 0\}$ are listed across the bottom, and results are arranged according to how many outcomes the predictor is actually associated with. The figure shows that SMRT is uniformly more powerful than MRT, Bonferroni and Sup, with the differences becoming

more apparent in identifying multiple regulation.

The above describes the results for testing a single predictor x_j with associated hypotheses $\mathcal{H}_j$. Results for controlling the FWER across all predictors and all hypotheses were qualitatively similar. SMRT maintained the nominal level of the test and obtained higher power than MRT, Sup, and Bonferroni at all sample sizes.

1.6 Discussion

We have proposed a framework for testing and estimation across a diverse set of outcomes. This framework allows the combination of information across continuous, semi-continuous, and discrete outcomes while maintaining control of the FWER for each predictor or across all predictors. It is flexible to the type of marginal likelihood specified and can easily incorporate more complex data structures such as censored survival outcomes. We rely on sparse estimation via penalization in our testing procedure. Many penalty functions could potentially accomplish similar results to the hierarchical penalty we proposed. As long as sparsistency holds and a suitable reference distribution can be obtained through permutation or resampling, other penalty functions could be worth exploring.

For simplicity, we used a working independence assumption to combine the profile log-likelihoods of multiple outcomes. But when the outcomes are not independent, incorporating information about the covariance in $\mathbf{y}$ can improve efficiency (Liang and Zeger, 1986). A further advantage to using the quadratic approximation to $\mathcal{L}^{(m)}$ in (1.4) instead of the profile log-likelihood itself (besides computational tractability) is that we can incorporate covariance information about $\mathbf{y}$ through the initial estimate $\widetilde{\boldsymbol{\beta}}$. If the (unpenalized) initial estimate $\widetilde{\boldsymbol{\beta}}$ is estimated in a way that gains efficiency by taking correlation in $\mathbf{y}$ into account, then that increase in efficiency will be propagated into our estimation of $\widehat{\boldsymbol{\beta}}$.

Finally, we have focused on FWER as the error rate of primary interest throughout this paper, but we could easily extend our framework to include more generalized error rates, such as k-FWER or the false discovery proportion, as in (Romano et al., 2010). We focus on FWER here because inference on a single predictor x_j when the number of outcomes M is not large likely calls for control of FWER. However, in some situations – especially when

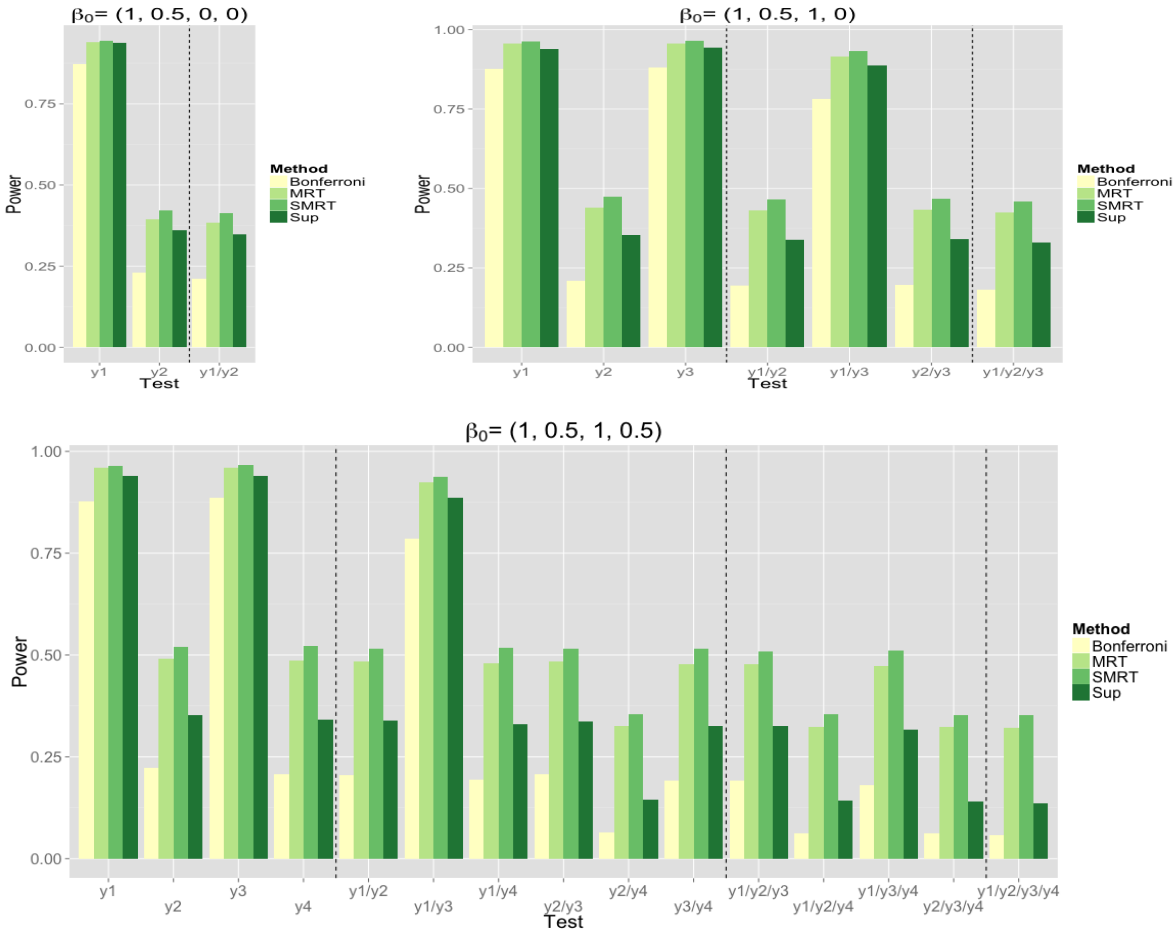


Figure 1.7: Power to detect non-null effects across 1000 simulations at sample size $n = 250$. Each plot indicates how many outcomes the predictors tested are associated with. For example, the top left plot corresponds to predictors with strong association to $y^{(1)}$, $\beta_{0j}^{(1)} = 1$ and weak association to $y^{(2)}$, $\beta_{0j}^{(2)} = 0.5$. Tests are listed on the x-axis. Power is indicated on the y-axis. Power estimates are aggregated over all estimates that share the same effect sizes. To take a couple of examples, the bar corresponding to "$y1$" in the figure corresponds to power to reject $H_j^{(1)}$, and the bar corresponding to "$y1/y2/y3$" in the figure corresponds to power to reject each of $H_j^{(1)}, H_j^{(2)}, H_j^{(3)}$ simultaneously.

the total number of tests T is large – one may desire to control a less restrictive error rate.

1.7 Appendix

For the following proofs, we put mild restrictions on the model (1.1) for each outcome $y^{(m)}$, as described in section 3 of Murphy and Van der Vaart (2000). We reproduce the restrictions here for completeness. Since the requirements hold for each outcome, we drop the superscripts $^{(m)}$ for the moment. Let $\log l(\boldsymbol{\beta}, h)(x)$ be the full log-likelihood. We require that there exists a $h_t(\boldsymbol{\beta}, h)$ such that $\ell(t, \boldsymbol{\beta}, h)(x) = \log l(t, h_t(\boldsymbol{\beta}, h))(x)$ is twice continuously differentiable for all x with first and second derivatives denoted $\dot{\ell}(t, \boldsymbol{\beta}, h)(x)$ and $\ddot{\ell}(t, \boldsymbol{\beta}, h)(x)$. Further, $h_{\boldsymbol{\beta}}(\boldsymbol{\beta}, h) = h$, for every $(\boldsymbol{\beta}, h)$. And $\dot{\ell}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0, h_0)$ must be the efficient score function. For every fixed $\boldsymbol{\beta}$, let $\widehat{h}_{\boldsymbol{\beta}}$ be the NPMLE for h. Then, for any $\boldsymbol{\beta}^\dagger \to_p \boldsymbol{\beta}_0$, $\widehat{h}_{\boldsymbol{\beta}^\dagger} \to_p h_0$ and $E\left[\dot{\ell}(\boldsymbol{\beta}_0, \boldsymbol{\beta}^\dagger, \widehat{h}_{\boldsymbol{\beta}^\dagger})\right] = o_p(\|\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}_0\| + n^{-1/2})$. Finally, suppose that there exists a neighborhood $\mathcal{W}$ of $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0, h)$ such that $\{\dot{\ell}(t, \boldsymbol{\beta}, h) : (t, \boldsymbol{\beta}, h) \in \mathcal{W}\}$ is Donsker with square integrable envelope function and $\{\ddot{\ell}(t, \boldsymbol{\beta}, h) : (t, \boldsymbol{\beta}, h) \in \mathcal{W}\}$ is Glivenko-Cantelli and bounded in L_1.

1.7.1 Justification of stepdown procedure

In this section, we will show that, when using an estimator that satisfies $P(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$, the testing procedure for each of the set of hypotheses $\{\mathcal{H}_j\}_{j=1,..,p}$ delineated in section 1.2 has FWER converging to 0 for any choice of $\psi \in [0, 1]$.

First we show that the stepdown procedure for $\mathcal{H}_j$ has FWER converging to 0. Goeman and Solari (2010) show that two conditions need to be satisfied in order for a sequentially rejective procedure of this sort to control the FWER at a given level α. First, a *monoticity* condition requires that the threshold for rejection must not increase as the test proceeds. That is, for $\Omega_{k'} \subseteq \Omega_k$:

$$c_j^{\Omega_k}(\psi) \geq c_j^{\Omega_{k'}}(\psi). \tag{1.5}$$

This condition is guaranteed by construction. Consider for any j, any element of $S_j^{*\Omega_k}$, $m_j^{*\Omega_k} \equiv \max_{m \in \Omega_k} t_j^{*(m)} \in S_{K_j}$ and the corresponding element of $S_j^{\Omega_{k'}}$, $m_j^{*\Omega_{k'}}$. Since $\Omega_{k'} \subset \Omega_k$,

$$m_j^{*\Omega_k} \geq m_j^{*\Omega_{k'}},$$

which in turn implies (1.5). Second, a *single-step* condition requires that the thresholds must be chosen so as to control type I error at α in the *critical case*, when the set of candidate hypotheses are all null. That is, recalling that $\mathscr{R}_{0j}$ is the set of indices of all true null hypotheses, $P\left(S_j^{\mathscr{R}_{0j}} > c_j^{\mathscr{R}_{0j}}(\psi)\right) \leq \alpha$. Because $P(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0) \to 1$, any choice of ψ will be sufficient. That is, for any ψ, $P\left(s_j^{\mathscr{R}_{0j}} > c_j^{\mathscr{R}_{0j}}(\psi)\right) \to 0$ because $P(s_j^{\mathscr{R}_{0j}} = 0) \to 1$ and $c_j^{\mathscr{R}_{0j}}(\psi) \geq 0$ for any ψ. Thus, our testing procedure will control the FWER for each predictor x_j asymptotically at any level α for any choice of ψ. Since we can choose α as small as we want, the FWER for the set of hypotheses $\mathcal{H}_j$ converges to 0.

1.7.2 Proof of sparsistency and asymptotic normality

To state and prove the results, we will need some preliminaries. Our objective function can be written equivalently as solely a function of $\boldsymbol{\beta}$ rather than as a function of $\alpha_j^{(m)}$ and d_j, as shown in Theorem 1 of Zhou and Zhu (2010). For a fixed number of outcomes and fixed number of predictors, the objective function can be written

$$Q(\boldsymbol{\beta}) = \|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2 + n\lambda_n \sum_{j=1}^{p} \Big\{ \sum_{m=1}^{M} w_j^{(m)} |\beta_j^{(m)}| \Big\}^{1/2} \tag{1.6}$$

where $n\lambda_n = \sqrt{\lambda}$.

We first show the root-n consistency of our estimator $\widehat{\boldsymbol{\beta}}$ in the following lemma, which is used in the proof of sparsistency and asymptotic normality.

Lemma 1. (Root-n consistency) *For PLLs $\{\mathcal{L}^{(m)}(\boldsymbol{\beta}^{(m)})\}_{m=1,\ldots,M}$ that satisfy the regularity conditions listed above, if $\lambda_n = O_p(n^{-1/2})$, then there exists a local maximizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.*

Proof. Let $Q(\boldsymbol{\beta}) = \sum_{m=1}^{M} (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal} \widetilde{\mathbf{I}}^{(m)} (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)}) + p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta})$. We will show that for a given $\tau > 0$, $c = \min_{j,m}\{|\beta_{0j}^{(m)}| : \beta_{0j}^{(m)} \neq 0\}$ there exists a constant C such that

$P[\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\beta}_0)] \geq 1 - \tau$. Now consider

$$\begin{aligned} D(\mathbf{u}) = Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\beta}_0) &= \sum_{m=1}^{M} (\boldsymbol{\beta}_0^{(m)} + n^{-1/2}\mathbf{u} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal} \widetilde{\mathbf{I}}^{(m)} (\boldsymbol{\beta}_0^{(m)} + n^{-1/2}\mathbf{u} - \widetilde{\boldsymbol{\beta}}^{(m)}) \\ &\quad - \sum_{m=1}^{M} (\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal} \widetilde{\mathbf{I}}^{(m)} (\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)}) + n \left(p_{\lambda_n, \mathbf{w}}(|\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}|) - p_{\lambda_n, \mathbf{w}}(|\boldsymbol{\beta}_0|) \right) \\ &= n^{-1/2} \mathbf{u}^{\intercal} \widetilde{\mathbf{I}} (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}) + \frac{n^{-1}}{2} \mathbf{u}^{\intercal} \widetilde{\mathbf{I}} \mathbf{u} - n \left(p_{\lambda_n, \mathbf{w}}(|\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}|) + p_{\lambda_n, \mathbf{w}}(|\boldsymbol{\beta}_0|) \right) \\ &= (I) + (II) + (III) \end{aligned}$$

Now, since $\|\widetilde{\mathbf{I}}^{(m)} - n\mathbf{I}^{(m)}\| = o_p(1)$, then $\|\widetilde{\mathbf{I}} - n\mathbf{I}\| = o_p(1)$, and $(I) = n^{\frac{1}{2}} \mathbf{u}^{\intercal} \mathbf{I} (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}})[1 + o_p(1)] \leq O_p(1)\|\mathbf{u}\|\|\mathbf{I}\|$. Furthermore, $(II) = \mathbf{u}^{\intercal}\mathbf{I}\mathbf{u}[1 + o_p(1)] \leq O_p(1)\|\mathbf{u}\|^2\|\mathbf{I}\|$. Now, following the argument in Zhou and Zhu (2010), $(III) \leq O_p(\lambda_n n^{\frac{1}{2}})$. Thus, as long as $\lambda_n = O_p(n^{-1/2})$, all terms are dominated by the first term of (II), which is positive. And the proof is concluded. □

We will now show that $\widehat{\boldsymbol{\beta}}$ is sparsistent: $P\left(\widehat{\beta}_{\mathcal{A}^c} = \mathbf{0}\right) \to 1$. If we can show that $\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j^{(m)}} = O_p(n^{\frac{1}{2}}) + n\frac{\partial p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta})}{\partial \beta_j^{(m)}}$ then sparsistency follows from lemma 1 and the argument in the proof of Theorem 4 in Zhou and Zhu (2010). To this end, note that

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j^{(m)}} &= (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal} \widetilde{\mathbf{I}}_j^{(m)} + n\frac{\partial p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta})}{\partial \beta_j^{(m)}} \\ &= (\boldsymbol{\beta}_0^{(m)} - \widetilde{\boldsymbol{\beta}}^{(m)})^{\intercal} \widetilde{\mathbf{I}}_j^{(m)} + (\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}_0^{(m)}) \widetilde{\mathbf{I}}_j^{(m)} + n\frac{\partial p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta})}{\partial \beta_j^{(m)}} = O_p(n^{\frac{1}{2}}) + n\frac{\partial p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta})}{\partial \beta_j^{(m)}} \end{aligned}$$

for any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$, noting that $\widetilde{\mathbf{I}}_j^{(m)} = O_p(n)$ for all j, m. And thus sparsistiency $P(\widehat{\beta}_{\mathcal{A}^c} = \mathbf{0}) \to 1$ follows.

Next we consider asymptotic normality. Let $\boldsymbol{\beta}(\mathcal{A})$ denote $\boldsymbol{\beta}$ with elements not in $\mathcal{A}$ set to 0. Because we have sparsistency, $\widehat{\boldsymbol{\beta}}(\mathcal{A})$ is a root-n consistent minimizer of $Q(\boldsymbol{\beta})$, and $\nabla Q\{\widehat{\boldsymbol{\beta}}(\mathcal{A})\} = o_p(1)$. Thus, minimizing $Q\{\boldsymbol{\beta}(\mathcal{A})\}$ is asymptotically equivalent to minimizing $Q_{\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}}) = (\boldsymbol{\beta}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}})^{\intercal} \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}} (\boldsymbol{\beta}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}) - (\boldsymbol{\beta}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}})^{\intercal} \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}^c} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^c} + n p_{\lambda_n, \mathbf{w}}(\boldsymbol{\beta}_{\mathcal{A}})]$ where $\widetilde{\mathbf{I}}_{\Omega_1, \Omega_2}$ denotes the submatrix of $\widetilde{\mathbf{I}}$ corresponding to rows in Ω_1 and columns in Ω_2. It follows that

$$\begin{aligned} o_p(1) = \nabla Q_{\mathcal{A}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) &= \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}) - \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}^c} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^c} + \nabla n p_{\lambda_n, \mathbf{w}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) \\ &= \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) + \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}} (\boldsymbol{\beta}_{0\mathcal{A}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}}) - \widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}^c} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^c} + \nabla n p_{\lambda_n, \mathbf{w}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) \end{aligned}$$

and hence

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) = n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) - n^{\frac{1}{2}}\widetilde{\mathbf{I}}^{-1}_{\mathcal{A},\mathcal{A}}\widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}^c}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^c} + n^{\frac{1}{2}}(n\widetilde{\mathbf{I}}^{-1}_{\mathcal{A},\mathcal{A}})\nabla p_{\lambda_n,\mathbf{w}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})$$
$$= (n\widetilde{\mathbf{I}}^{-1}_{\mathcal{A},\mathcal{A}})n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0) + n^{\frac{1}{2}}(n\widetilde{\mathbf{I}}^{-1}_{\mathcal{A},\mathcal{A}})\nabla p_{\lambda_n,\mathbf{w}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) + o_p(1)$$

This, together with the same argument as in the proof of Theorem 4 in Zhou and Zhu (2010), $\nabla p_{\lambda_n,\mathbf{w}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) = o_p(n^{-\frac{1}{2}})$, implies that $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) = n^{-1/2}\mathbf{I}^{-1}_{\mathcal{A},\mathcal{A}}\sum_{i=1}^{n}\boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0) + o_p(1)$.

1.7.3 Properties of resampled $\widehat{\boldsymbol{\beta}}^*$

Let $\mathbb{P}^*$ be the measure generated by both $\mathbb{V}$ and $\mathcal{G}$. First, note that

$$\left\|\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\right\| = \left\|\widetilde{\boldsymbol{\beta}} + \sum_{i=1}^{n}\widetilde{\mathbf{I}}^{-1}\widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})(G_i - 1) - \boldsymbol{\beta}_0\right\| \leq \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| + \left\|\sum_{i=1}^{n}\widetilde{\mathbf{I}}^{-1}\widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})(G_i - 1)\right\|$$
$$= O_{\mathbb{P}^*}(n^{-1/2}) + \left\|\frac{1}{n}\sum_{i=1}^{n}\left\{n\widetilde{\mathbf{I}}^{-1}\widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})\right\}(G_i - 1)\right\|$$

Noting that $\mathcal{G}$ is independent of $\mathbb{V}$, $E[G_i - 1] = 0$, and $E[\mathbf{I}^{-1}\widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})] < \infty$, $\frac{1}{n}\sum_{i=1}^{n}\{n\widetilde{\mathbf{I}}^{-1}\widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}})\}(G_i - 1) \to_{\mathbb{P}^*} 0$ and the perturbed initial estimate is also root-n consistent: $\left\|\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\right\| = O_{\mathbb{P}^*}(n^{-1/2})$.

Moreover, for the root-n consistency proof of $\widehat{\boldsymbol{\beta}}$, the role of $\widetilde{\boldsymbol{\beta}}$ in $Q(\boldsymbol{\beta})$ is only that of a root-n consistent initial estimate. Inspection of the proof of $\widehat{\boldsymbol{\beta}}$'s root-$n$ consistency will show that the only fact about $\widetilde{\boldsymbol{\beta}}$ that we need is $\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| = O_p(n^{-1/2})$. Therefore, in just the same way, root-n consistency of $\widetilde{\boldsymbol{\beta}}^*$ gives us root-n consistency of $\widehat{\boldsymbol{\beta}}^*$: $\left\|\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\right\| = O_{\mathbb{P}^*}(n^{-1/2})$. Now, sparsistency of $\widehat{\boldsymbol{\beta}}^* \mid \mathbb{V}$ follows from a similar argument as for sparsistency of $\widehat{\boldsymbol{\beta}}$. Consider

$$\left.\frac{\partial Q^*(\boldsymbol{\beta})}{\partial\beta_j^{(m)}}\right|\mathbb{V} = (\boldsymbol{\beta}^{(m)} - \widetilde{\boldsymbol{\beta}}^{*(m)})^{\intercal}\widetilde{\mathbf{I}}_j^{(m)} + n\left.\frac{\partial p_{\lambda_n,\mathbf{w}^*}(\boldsymbol{\beta})}{\partial\beta_j^{(m)}}\right|\mathbb{V} = O_{\mathbb{P}^*}(n^{\frac{1}{2}}) + n\left.\frac{\partial p_{\lambda_n,\mathbf{w}}(\boldsymbol{\beta})}{\partial\beta_j^{(m)}}\right|\mathbb{V}$$

for any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_{\mathbb{P}^*}(n^{-1/2})$, noting that $\sum_{i=1}^{n}(G_i - 1) = O_{\mathbb{P}^*}(n^{\frac{1}{2}})$. And thus sparsistency follows: $P\left(\widehat{\beta}^*_{\mathcal{A}^c} = \mathbf{0}\middle|\,\mathbb{V}\right) \to 1$.

Finally, following the logic in the proof of asymptotic normality of $\widehat{\boldsymbol{\beta}}$,

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}^*_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) = n^{\frac{1}{2}}(\widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) - n^{\frac{1}{2}}\widetilde{\mathbf{I}}^{-1}_{\mathcal{A},\mathcal{A}}\widetilde{\mathbf{I}}_{\mathcal{A},\mathcal{A}^c}\widetilde{\boldsymbol{\beta}}^*_{\mathcal{A}^c} + o_{\mathbb{P}^*}(1) = n^{\frac{1}{2}}\mathbf{K}(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) + o_{\mathbb{P}^*}(1)$$

for $\mathbf{K} = d_{\mathcal{A}} - \widetilde{\mathbf{I}}_{\mathcal{A}\mathcal{A}}^{-1}\widetilde{\mathbf{I}}_{\mathcal{A}\mathcal{A}^c}d_{\mathcal{A}^c}$, $d_{\mathcal{A}}\boldsymbol{\beta} = \boldsymbol{\beta}_{\mathcal{A}}$, and $d_{\mathcal{A}^c}\boldsymbol{\beta} = \boldsymbol{\beta}_{\mathcal{A}^c}$. In the proof of asymptotic normality of $\widehat{\boldsymbol{\beta}}$, we showed that $n^{1/2}\mathbf{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{-1/2}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1}\sum_{i=1}^n \boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0) + o_{\mathbb{P}^*}(1)$. Note that $n^{\frac{1}{2}}\mathbf{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = n^{\frac{1}{2}}\mathbf{K}\widetilde{\mathbf{I}}^{-1}\sum_{i=1}^n \widetilde{\boldsymbol{\varphi}}_i(\widetilde{\boldsymbol{\beta}}) + o_{\mathbb{P}^*}(1)$, which suggests

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^* - \boldsymbol{\beta}_{0\mathcal{A}}) = n^{-1/2}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1}\sum_{i=1}^n \boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0)(G_i - 1) + n^{\frac{1}{2}}\mathbf{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_{\mathbb{P}^*}(1)$$

And recall from above that $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) = n^{-1/2}\mathbf{I}_{\mathcal{A},\mathcal{A}}^{-1}\sum_{i=1}^n \boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0) + o_{\mathbb{P}^*}(1) = n^{\frac{1}{2}}\mathbf{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_{\mathbb{P}^*}(1)$. Then, let $\mathbf{Z}_i = n^{-1/2}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\varphi}_{i\mathcal{A}}(\boldsymbol{\beta}_0)(G_i - 1)$, so that $E[\mathbf{Z}_i|\mathbb{V}] = 0$ and $\mathrm{Cov}[\mathbf{Z}_i|\mathbb{V}] = n^{-1}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\varphi}_{i\mathcal{A}}\boldsymbol{\varphi}_{i\mathcal{A}}^{\intercal}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1} \equiv \Gamma_i$. Because $\sum_{i=1}^n E[\|\Gamma_i^{-1/2}\|_2^3|\mathbb{V}] = o_{\mathbb{P}^*}(1)$, then by the argument in Bentkus (2005), $n^{\frac{1}{2}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^* - \widehat{\boldsymbol{\beta}}_{\mathcal{A}})\Big|\,\mathbb{V} \to_{\mathcal{L}} N(0, \mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\mathbf{I}_{\mathcal{A}\mathcal{A}}^{-1})$.

1.7.4 Algorithm

An iterative procedure can be employed to fit the model (1.4). First, fix $\mathbf{d}$ and estimate $\boldsymbol{\alpha}$ via adaptive lasso. Next, fix $\boldsymbol{\alpha}$ and estimate $\mathbf{d}$ using the nonnegative garrote. However, because of the widespread availability and speed of lasso-type estimation, we in general prefer to employ adaptive lasso to the nonnegative garrote. So we propose to estimate $\mathbf{d}$ using adaptive lasso as well, by minimizing the following objective function

$$\|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^p |d_j| + \lambda\sum_{m=1}^M\sum_{j=1}^p w_j^{(m)}|\alpha_j^{(m)}|, \tag{1.7}$$

Using the adaptive lasso in place of the nonnegative garrote is justified here by the argument in Zou (2006) that the adaptive lasso is asymptotically equivalent to the nonnegative garrote. That is, the nonnegative garrote is equivalent to the adaptive lasso with a further sign constraint, and the sign constraint is satisfied (at least in the limit) by consistency of the adaptive lasso. Our iterative fitting procedure, then, uses the adaptive lasso at both stages and is thus very fast.

Fitting the model can proceed as follows:

1. Set $\mathbf{d}_{(0)} = \mathbf{1}$ and $\widetilde{\mathbb{X}}_\beta = \widetilde{\mathbb{X}}\mathrm{diag}(|\widetilde{\boldsymbol{\beta}}|)$. Let $k = 1$.
2. Update $\boldsymbol{\alpha}$. Set $D_\alpha = \mathrm{diag}(\mathbf{d}_{(k-1)})$ and $\widetilde{\mathbb{X}}_\alpha = \widetilde{\mathbb{X}}_\beta\mathrm{diag}(D_\alpha, ..., D_\alpha)_{Mp\times Mp}$ and obtain

$$\boldsymbol{\alpha}_{(k)} = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\,\|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}_\alpha\boldsymbol{\alpha}\|_2^2 + \lambda\sum_{m=1}^M\sum_{j=1}^p |\alpha_j^{(m)}|$$

Inclusion of $\widetilde{\boldsymbol{\beta}}$ in $\widetilde{\mathbb{X}}_\alpha$ is equivalent to using weights $w_j^{(m)} = |\widetilde{\beta}_j^{(m)}|^{-1}$.

3. Update $\mathbf{d}$. Set $\widetilde{\mathbb{X}}_d = \widetilde{\mathbb{X}}\mathbf{A}_d$ where $\mathbf{A}_d = \begin{bmatrix} \text{diag}(\boldsymbol{\alpha}_{(k)}^{(1)})_{p\times p} \\ \text{diag}(\boldsymbol{\alpha}_{(k)}^{(2)})_{p\times p} \\ \vdots \\ \text{diag}(\boldsymbol{\alpha}_{(k)}^{(M)})_{p\times p} \end{bmatrix}_{Mp\times p}$

Then,

$$\mathbf{d}_{(k)} = \underset{\mathbf{d}}{\text{argmin}} \, \|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}_d \mathbf{d}\|_2^2 + \sum_{j=1}^{p} |d_j|$$

4. Update $\boldsymbol{\beta}$.

$$\beta_{j(k)}^{(m)} = d_{j(k)} \alpha_{j(k)}^{(m)} |\widetilde{\beta}_j^{(m)}|$$

.

5. Iterate until convergence.

Semiparametric canonical correlation analysis

Denis Agniel and Tianxi Cai
Department of Biostatistics
Harvard School of Public Health

2.1 Introduction

In recent years, considerable interest has been focused on studying multiple phenotypes simultaneously in both epidemiological and genomic studies. Such studies are important for several reasons. First, a complex disorder is usually associated with multiple correlated phenotypes. Hence, even when a study is focused on a specific disease, multiple phenotypes might be needed to fully capture the complexity and multidimensionality of the disorder. Second, multiple related disorders might share the same etiology and a joint assessment will enable identifying subpopulations which might be at high risk of multiple diseases.

For example, hypertension, type 2 diabetes, and cardiovascular diseases are common chronic conditions that tend to occur in the same individual, suggesting common predisposing factors (Cheung, 2010). Established common environmental factors that predispose to these diseases include obesity and physical activity. Individuals with abdominal obesity are likely to develop lipid abnormalities and elevation of blood pressure and glucose. In time, hypertension, diabetes and cardiovascular disease ensue. For psychiatric disorders, many diseases also are associated with childhood adversities, suggesting shared environmental roots (Kessler et al., 1997; Green et al., 2010). Recent genetic studies have suggested that many genetic loci appear to harbor variants associated with multiple traits (Solovieff et al., 2013). For example, genetic epidemiological studies have documented that schizophrenia, bipolar disorder, and major depressive disorder share familial and genetic determinants (Smoller et al., 2013). As another example, recent studies have identified common genes associated with a higher risk of what were previously considered distinct autoimmune diseases (Zhernakova et al., 2009; Xavier and Rioux, 2008).

When multiple related disorders share common genetic and/or environmental factors, one may question whether these disorders are etiologically distinct. If these disorders do have a shared etiology, it is possible to identify a new common underlying trait and examine how well known risk factors can explain the underlying trait. If a risk profile can be developed to accurately predict such a trait, prevention and treatment strategies can then be developed to better manage the disorders.

A useful approach to address such questions is the canonical correlation analysis (CCA)

(Hotelling, 1936; Thompson, 1984; Hardoon et al., 2004). In CCA, traits and risk profiles are identified as $\boldsymbol{a}_1^\intercal \mathbf{y}$ and $\mathbf{b}_1^\intercal \mathbf{x}$, respectively, via the maximization problem

$$\underset{\boldsymbol{a},\mathbf{b}}{\operatorname{argmax}} \operatorname{cor}(\boldsymbol{a}^\intercal \mathbf{y}, \mathbf{b}^\intercal \mathbf{x}), \tag{2.1}$$

for vector of phenotypes $\mathbf{y} = (y_m)_{1 \le m \le M}$ and vector of predictors $\mathbf{x} = (x_j)_{1 \le j \le p}$ subject to a constraint on the norms of $\boldsymbol{a}$ and $\mathbf{b}$ for identifiability. Additional projections pairs $(\boldsymbol{a}_k, \mathbf{b}_k)_{k>1}$, orthogonal to $(\boldsymbol{a}_1, \mathbf{b}_1)$ and each other, can also be sequentially identified to capture residual correlations between the outcomes and predictors. However, since the phenotypes quantifying different disorders are often of different scales, identifying traits as linear combinations of $\mathbf{y}$ may not be appropriate or possible. For example, when the outcomes are measurements of biomarkers for one autoimmune disease, the measurements of one biomarker may be orders of magnitude larger or smaller than the measurements for another biomarker. Hence, transformations of the outcomes are needed in order to put them onto the same scale and relate them to the predictors. Furthermore, when some outcomes are subject to censoring or truncation, the traditional CCA is also not applicable since the relevant covariances are not identifiable without modeling. As an example, the biomarkers for some disorders like autoimmune disease may be subject to limits of quantification above and below which the true measurement of the biomarkers are unavailable.

Various extensions of CCA have been proposed in recent years. Other measures of association beyond Pearson's correlation coefficient (Jin and Cui, 2010) and semiparametric single-index models (Xia, 2008) have been proposed to account for nonlinearity between the traits and risk profiles, but these methods still require that traits be linear combinations of $\mathbf{y}$. In the setting of fully observed, continuous $\mathbf{y}$, (Zhu et al., 2007) proposes to estimate transformations of $\mathbf{y}$ via smoothing and goes on to perform partial least squares, a close relative of CCA. No methods have yet been proposed for our situation where $\mathbf{y}$ may be *diverse* – that is, $\mathbf{y}$ may contain components that are continuous, discrete, and/or not fully observed due to censoring or truncation.

We propose *semiparametric canonical correlation analysis* (sCCA) to identify projection pairs $(\boldsymbol{a}_k, \mathbf{b}_k)_{k \ge 1}$ where now traits are conceived as $\boldsymbol{a}_k^\intercal \mathbf{h}$ where $\mathbf{h}$ is a vector of transformed (possibly latent) phenotypes, all of which are on the same scale. We identify the projection

pairs via a three-stage procedure. The first stage involves marginal models that put all phenotypes on the same scale by conceiving of each y_m as a possibly thresholded or truncated observation of a smooth latent variable z_m. We use semiparametric transformation models for continuous or censored y_m and parametric models for discrete y_m. In the second stage, we characterize the covariance matrix of $\mathbf{h}$ using copula methods in conjunction with the estimates from the marginal models. Copulas have been used in similar modeling situations for survival data (Othus and Li, 2010) and discrete data (Xue-Kun Song, 2000). And finally we perform CCA using the results from the copula and marginal models to identify the sCCA projection pairs. Note that the target of estimation is the set of projection pairs, not the traits themselves. By obtaining the projection pairs, we can obtain the risk profiles $(\mathbf{b}_k^\intercal \mathbf{x})_{k\geq 1}$, but since $\mathbf{h}$ may involve the unobservable z_m, we may not recover the traits $(\boldsymbol{a}_k^\intercal \mathbf{h})_{k\geq 1}$ even when we can recover the trait directions $(\boldsymbol{a}_k)_{k\geq 1}$.

The structure of the rest of the paper is as follows. In section 2.2, we introduce sCCA and discuss its properties. In section 2.3, we apply sCCA to a genetic study to identify risk profiles for autoimmune disease, and we provide simulation results. Finally, in section 2.4, we provide a discussion of the method.

2.2 Semiparametric canonical correlation analysis

Suppose the data for analysis consists of n independent and identically distributed random vectors $\mathbb{V} = \{\mathbf{V}_i = (\mathbf{y}_i^\intercal, \boldsymbol{\delta}_i^\intercal, \mathbf{x}_i^\intercal)^\intercal\}_{i=1,...,n}$ where $\mathbf{y}_i = (y_{i1}, ..., y_{iM})^\intercal$ is a set of M phenotypes, $\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{iM})^\intercal$ is a set of M censoring indicators, and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^\intercal$ is a set of p predictors for the the ith subject.

Recall that, if we were interested in traits of the form $\boldsymbol{a}_k^\intercal \mathbf{y}$, we would seek to obtain (2.1), or equivalently,

$$\underset{\boldsymbol{a},\mathbf{b}}{\operatorname{argmax}}\, \boldsymbol{a}^\intercal \Sigma_{yx} \mathbf{b}, \qquad \text{subject to } \boldsymbol{a}^\intercal \Sigma_{yy} \boldsymbol{a} = \mathbf{b}^\intercal \Sigma_{xx} \mathbf{b} = 1$$

where $\Sigma_{wv} = \operatorname{Cov}(\mathbf{w}, \mathbf{v})$. However, due to the components of $\mathbf{y}$ being incompletely observed or measured on incomparable scales, identifying traits as linear combinations of $\mathbf{y}$ does not

make sense. Instead, we consider traits to be of the form $\boldsymbol{a}_k^\intercal \mathbf{h}$ and we thus obtain

$$(\boldsymbol{a}_1, \mathbf{b}_1) = \underset{\boldsymbol{a},\mathbf{b}}{\operatorname{argmax}}\, \boldsymbol{a}^\intercal \Sigma_{hx} \mathbf{b}, \qquad \text{subject to } \boldsymbol{a}^\intercal \Sigma_{hh} \boldsymbol{a} = \mathbf{b}^\intercal \Sigma_{xx} \mathbf{b} = 1. \tag{2.2}$$

Additional directions $(\boldsymbol{a}_k, \mathbf{b}_k)$ can be obtained by solving (2.2) with the additional constraints that $(\boldsymbol{a}_k^\intercal \boldsymbol{a}_j = \mathbf{b}_k^\intercal \mathbf{b}_j = 0)_{1 \le j < k}$. Another quantity of interest is the *proportion of variance explained* $r_k = \operatorname{tr}(\Sigma_{hh} \boldsymbol{a}_k \rho_k^2 \boldsymbol{a}_k^\intercal \Sigma_{hh}) / \operatorname{tr}(\Sigma_{hh})$ (Glahn, 1968) where $\rho_k = \operatorname{cor}(\boldsymbol{a}_k^\intercal \mathbf{h}, \mathbf{b}_k^\intercal \mathbf{x})$.

Thus the parameters of interest, $\boldsymbol{\theta} = (\boldsymbol{a}_k, \mathbf{b}_k, r_k)_{1 \le k \le K}$ can be expressed as functions of $\boldsymbol{\sigma} = (\operatorname{vec}(\Sigma_{xx})^\intercal, \operatorname{vec}(\Sigma_{hx})^\intercal, \operatorname{vec}(\Sigma_{hh})^\intercal)^\intercal$

$$\boldsymbol{\theta} = g(\boldsymbol{\sigma}). \tag{2.3}$$

And we estimate $g(\boldsymbol{\sigma})$ by simply plugging in suitable consistent estimators

$$\widehat{\boldsymbol{\theta}} = g(\widehat{\boldsymbol{\sigma}}).$$

In section 2.2.1, we identify $\mathbf{h}$, Σ_{hh}, and Σ_{hx}, and in section 2.2.2 we take up estimation.

2.2.1 Identifying h

Since the diversity of $\mathbf{y}$ causes its components to be on incomparable scales and possibly unobserved, we first identify possibly latent continuous variables $\mathbf{z} = (z_m)_{1 \le m \le M}$ whose components can be put on comparable scales. The relationship between the observed y_m and the latent z_m can be described as follows. If y_{im} is continuous and uncensored ($\delta_{im} = 0$), then $y_{im} = z_{im}$. If y_{im} is continuous and censored ($\delta_{im} = 1$), then $y_{im} \le z_{im}$. For discrete outcomes, we consider y_m to be a thresholded version of z_m such that, if y_m, without loss of generality, takes values in the set $\mathcal{Y}_m = \{0, 1, ..., K_m\}$ then

$$y_{im} = \begin{cases} 0, & z_{im} \le 0 \\ k, & z_{im} \in (k-1, k], \qquad 1 \le k \le K_m - 1 \\ K_m, & z_{im} > K_m - 1. \end{cases}$$

We assume that the marginal distribution of z_m follows the following model

$$P(z_m \le z \mid \mathbf{x}) = g_m\{\mathbf{x}^\intercal \boldsymbol{\beta}_{m0} + h_{m0}(z)\}, \tag{2.4}$$

where $\boldsymbol{\beta}_{m0}$ represents the unknown effect of $\mathbf{x}$ on z_m and the link function, $g_m(\cdot)$, is given although the correlation structure of $\mathbf{z}$ (and thus $\mathbf{y}$) is left unspecified. The transformation $h_{m0}(\cdot)$ is an infinite-dimensional, unspecified, smooth, increasing function if y_m is continuous, while it is a finite-dimensional function on the set $\{0, ..., K_m-1\}$ if y_m is discrete. Choice of g_m determines the type of model being fit. Some practical examples include $g_m(x) = e^x/(1+e^x)$, corresponding to a proportional odds model for continuous y_m and a logistic regression model if y_m is binary. One may let $g_m(x) = 1 - e^{-e^x}$ to impose a proportional hazards model.

The model (2.4) is equivalent to

$$h_{m0}(z_m) = -\mathbf{x}^\intercal \boldsymbol{\beta}_{m0} + \epsilon_m, \quad \epsilon_m \sim g_m. \tag{2.5}$$

In this representation, it is clear that the components of $\mathbf{h} = (h_{m0}(z_m))_{1\leq m\leq M}^\intercal$ are on comparable scales, as each is expressed as a linear combination of $\mathbf{x}$ and an independent error. We define

$$\Sigma_{hh} = \mathrm{Cov}(\mathbf{h}, \mathbf{h}) = \mathbb{B}_0^\intercal \Sigma_{xx} \mathbb{B}_0 + \Sigma_{\epsilon\epsilon} \text{ and } \Sigma_{hx} = \mathrm{Cov}(\mathbf{h}, \mathbf{x}) = \Sigma_{xx}\mathbb{B}_0, \tag{2.6}$$

where $\mathbb{B}_0$ is a $K \times p$ matrix with columns $\boldsymbol{\beta}_{m0}$ and $\Sigma_{\epsilon\epsilon} = \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}), \boldsymbol{\epsilon} = (\epsilon_m)_{1\leq m\leq M}$.

2.2.2 Estimating the joint distribution

In order to estimate $\boldsymbol{\sigma}$, we must provide estimates of Σ_{xx}, Σ_{hx}, and Σ_{hh}. Estimation of Σ_{xx} may proceed in the usual fashion, using

$$\widehat{\Sigma}_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\intercal, \tag{2.7}$$

the regular unbiased estimator of Σ_{xx}. Furthermore, recalling that Σ_{hh} and Σ_{hx} depend on $\mathbb{B}_0$, we may obtain, $\widehat{\mathbb{B}}$, the $p \times K$ matrix with columns $\widehat{\boldsymbol{\beta}}_m$, where $\widehat{\boldsymbol{\beta}}_m$ is obtained by maximizing each marginal likelihood specified by (2.5), yielding $\widehat{\boldsymbol{\beta}}_m$ as the nonparametric maximum likelihood estimate (NPMLE)(Murphy and Van der Vaart, 2000) if (2.5) is semiparametric or as the maximum likelihood estimate (MLE) if (2.5) is parametric.

The only remaining quantity to estimate is $\Sigma_{\epsilon\epsilon}$. To estimate $\Sigma_{\epsilon\epsilon}$, we assume a copula model on the joint distribution of $\boldsymbol{\epsilon}$. We characterize the joint distribution function $G(\boldsymbol{\epsilon})$

through the copula

$$G(\boldsymbol{\epsilon}) = C_{\Sigma_{\epsilon\epsilon}}(G_1(\epsilon_1), ..., G_M(\epsilon_M)),$$

where $C_{\Sigma_{\epsilon\epsilon}}$ is a copula parametrized by the correlation matrix $\Sigma_{\epsilon\epsilon}$. By definition, for $\mathbf{u} = (u_m)_{1 \leq m \leq M}$, any function $C(\mathbf{u}) : (0,1)^M \to (0,1)$ is a copula if it is a continuous distribution function and

$$\lim_{u_j \to 1, j \neq m} C(\mathbf{u}) = u_m.$$

One practical example of a copula function is the Gaussian copula:

$$C_{\Sigma_{\epsilon\epsilon}}(\mathbf{u}) = \Phi_{\Sigma_{\epsilon\epsilon}}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_M)),$$

where $\Phi(\cdot)$ is the standard normal distribution function and $\Phi_{\Sigma_{\epsilon\epsilon}}(\cdot)$ is the distribution function of the centered multivariate normal with unit variances and correlation matrix $\Sigma_{\epsilon\epsilon}$. Recognizing that $\epsilon_m = \mathbf{x}^\intercal \boldsymbol{\beta}_{m0} + h_{m0}(z_m)$, we can write the distribution function for $\mathbf{z}$ conditional on $\mathbf{x}$ as

$$F(\mathbf{z}) = C_{\Sigma_{\epsilon\epsilon}}\left[G_1\left\{\mathbf{x}^\intercal \boldsymbol{\beta}_{10} + h_{10}(z_1)\right\}, ..., G_M\left\{\mathbf{x}^\intercal \boldsymbol{\beta}_{M0} + h_{M0}(z_M)\right\}\right].$$

Now, since $\mathbf{z}$ is not observed, observing the vector $\mathbf{y}$ and missing indicator $\boldsymbol{\delta}$ implies that $\mathbf{z}$ lies in some region, which in turn – conditional on $\mathbf{x}$ – implies that $\boldsymbol{\epsilon}$ lies in a particular region. For a toy example, suppose $\mathbf{y}^\intercal = (y_1, y_2, y_3) = (3, 5, 0)$ and $\boldsymbol{\delta}^\intercal = (0, 1, 0)$, where y_1 is continuous and fully observed, y_2 is continuous and censored, and y_3 is binary. That implies that $\mathbf{z}$ lies in the region $\{(z_1, z_2, z_3)^\intercal \in \mathbb{R}^3 : z_1 = 3, z_2 > 5, z_3 \leq 0\}$, which in turn implies that $\boldsymbol{\epsilon}$ lies in the region $\{(e_1, e_2, e_3)^\intercal \in \mathbb{R}^3 : e_1 = \mathbf{x}^\intercal \boldsymbol{\beta}_{10} + h^{(1)}(3), e_2 > \mathbf{x}^\intercal \boldsymbol{\beta}_{20} + h^{(2)}(5), e_3 \leq \mathbf{x}^\intercal \boldsymbol{\beta}_{30} + h(0)\}$.

Then the likelihood for $\Sigma_{\epsilon\epsilon}$ can clearly be written

$$\mathcal{L}(\Sigma_{\epsilon\epsilon}; \mathbb{V}, \mathcal{H}_0, \mathbb{B}_0) = \prod_{i=1}^{n} f(\mathbf{y}_i | \mathbf{x}_i, \Sigma_{\epsilon\epsilon}, \mathcal{H}_0, \mathbb{B}_0)$$

where $\mathcal{H}_0 = (h_{m0}(\cdot))_{1 \leq m \leq M}$ and $f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\delta}_i, \Sigma_{\epsilon\epsilon}, \mathcal{H}_0, \mathbb{B}_0)$ can be computed by integrating the copula density $c_{\Sigma_{\epsilon\epsilon}}\left[G_1\left\{\mathbf{x}^\intercal \boldsymbol{\beta}_{10} + h_{10}(z_1)\right\}, ..., G_M\left\{\mathbf{x}^\intercal \boldsymbol{\beta}_{M0} + h_{M0}(z_M)\right\}\right]$ over the region in which $\mathbf{z}_i$ lies.

We then obtain

$$\widehat{\Sigma}_{\epsilon\epsilon} = \underset{\Sigma_{\epsilon\epsilon}}{\operatorname{argmax}} \mathcal{L}(\Sigma_{\epsilon\epsilon}; \mathbb{V}, \widehat{\mathcal{H}}, \widehat{\mathbb{B}}) \tag{2.8}$$

where $\widehat{\mathcal{H}} = (\widehat{h}_m(\cdot))_{1\leq m\leq M}$ and $\widehat{h}_m(\cdot)$ is obtained as the NPMLE or MLE of $h_m(\cdot)$ from the marginal model (2.5).

Final estimates of $\boldsymbol{\sigma}$ are obtained from (2.7) and by plugging in (2.8), (2.7), and $\widehat{\mathbb{B}}$ into (2.6).

2.2.3 Assessing variability

In this section we examine the asymptotic distribution of our estimates $\widehat{\boldsymbol{\theta}}$. Recall that $\widehat{\boldsymbol{\theta}} = g(\widehat{\boldsymbol{\sigma}})$. Thus, taking a Taylor expansion,

$$\begin{aligned} n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= n^{\frac{1}{2}} \{g(\widehat{\boldsymbol{\sigma}}) - g(\boldsymbol{\sigma})\} \\ &= n^{\frac{1}{2}} \frac{\partial g(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} (\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}) + n^{\frac{1}{2}} O_p(\|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|^2) \end{aligned}$$

We show in the appendix that $n^{\frac{1}{2}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$ can be written as $n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{U}_{\sigma i}(\boldsymbol{\sigma})$ for some mean zero functions $\mathcal{U}_{\sigma i}(\boldsymbol{\sigma})$, and thus

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n^{-\frac{1}{2}} \frac{\partial g(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} \sum_{i=1}^{n} \mathcal{U}_{\sigma i}(\boldsymbol{\sigma}) + o_p(1)$$

which means that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal. The results in the appendix along with explicit forms of $\frac{\partial g(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}}$, which can be found in Ogasawara (2007), could be used to estimate the distribution of $\widehat{\boldsymbol{\theta}}$. However, the variance estimates based on this asymptotic result may be unstable and unlikely to perform well in finite samples. Thus, we employ the bootstrap to obtain variance estimates for $\widehat{\boldsymbol{\theta}}$.

Let $\{\widehat{\boldsymbol{\theta}}^*_b\}_{1\leq b\leq B}$ be a collection of B bootstrapped estimates of $\boldsymbol{\theta}$, where $\widehat{\boldsymbol{\theta}}^*_b = (\widehat{\theta}^*_{bj})_{1\leq j\leq K(M+p+1)} = (\widehat{\mathbf{a}}^*_{bk}, \widehat{\mathbf{b}}^*_{bk}, \widehat{r}^*_{bk})$. Then one could estimate the variability of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ by examining the variability of $n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_b - \widehat{\boldsymbol{\theta}})$. One can estimate the standard error of $\widehat{\theta}_j$ by computing the standard deviation of the collection of bootstrap values $\{\widehat{\theta}^*_{bj}\}_{1\leq b\leq B}$.

To construct a $(1-\alpha)\%$ confidence interval for θ_j, one can compute the so-called *percentile* interval as (C_l, C_u), where the endpoints are the lower and upper $\alpha/2$ quantiles of

$\{\widehat{\theta}^*_{bj}\}_{1\leq b\leq B}$, respectively. Because some of the estimators may be biased, it is also possible to compute an *adjusted percentile* interval, by adjusting the percentile interval according to an estimate of the bias. Let $\widehat{\text{bias}}_{\widehat{\theta}_j} = B^{-1}\sum_{b=1}^{B}\widehat{\theta}^*_{bj} - \widehat{\theta}_j$ be an estimate of the bias in $\widehat{\theta}_j$. Then the adjusted percentile interval for θ_j is computed as $(C_l - \widehat{\text{bias}}_{\widehat{\theta}_j}, C_u - \widehat{\text{bias}}_{\widehat{\theta}_j})$. Finally, one can compute a *basic* bootstrap interval as $(C_u - \widehat{\theta}_j, C_l - \widehat{\theta}_j)$.

2.2.4 Visualizing risk profiles

As described in section 2.2.2, a copula must be specified to obtain sCCA estimates. In practice, while there may be compelling scientific motivation for the structure of the marginal models (2.5), there will often be less clear scientific reasoning available for choice of the copula $C_{\Sigma_{\epsilon\epsilon}}(\cdot)$, and copulas may be specified solely for computational simplicity. Thus, one might be concerned about the dependence of sCCA estimates on a potentially misspecified copula.

To alleviate these concerns, one can examine the relationship between a risk profile $\mathbf{b}_k^{\mathrm{T}}\mathbf{x}$ and $\mathbf{y}$ nonparametrically. By smoothing, transformations of y_m over the risk profile $\mathbf{b}_k^{\mathrm{T}}\mathbf{x}$ one can get a clearer view of how the risk profile predicts values of y_m irrespective of the copula originally used.

2.3 Example

2.3.1 Genetic study to identify risk profiles for autoimmune disease

We apply our sCCA to a study of shared autoimmunity with the goal of identifying risk profiles associated with 4 autoantibodies: anti-nuclear antibodies (ANA), anti-cyclic citrullinated protein (CCP) antibodies, anti-transglutaminase (TTG) antibodies, and anti-thyroid peroxidase antibodies (TPO). These 4 autoantibodies are respectively markers for 4 autoimmune diseases (ADs): systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), celiac disease and autoimmune thyroid disease. The genetic markers consists of 67 single-nucleotide polymorphisms (SNPs) previously published as potential risk markers for these four ADs. We use the autoantibodies in place of the disease phenotypes because, while the co-occurrence of multiple ADs within individuals has been documented (Somers et al., 2006),

it would be rare, even for someone who is at high risk for the spectrum of ADs, to have more than one. In contrast, autoantibodies can be present in individuals predisposed to having the disease even in the absence of a disease phenotype. For example, while co-occurrence of ADs within families is well documented (Somers et al., 2006), family members of those with autoimmune disease may also experience elevated levels of autoantibodies if they haven't (yet) exhibited the disease phenotype. In this study, the autoantibodies are considered markers for subjects at higher risk for SLE, celiac, and autoimmune thyroid disease.

The study cohort includes 1265 individuals of European ancestry with RA identified through electronic medical records at Partners Healthcare (Liao et al., 2010; Kurreeman et al., 2011). Due to a limit of quantification, the antibody measurements are highly unreliable when the values are either very low or very high. ANA and TTG rarely reached the lower limit of quantification in our study cohort, with only 1 and 4 people below the limit, respectively. On the other hand, 28% of CCP measurements were below the limit, as well as 74% of TPO measurements. Similarly, few people were above the limit of quantification for ANA (42 people, 3%) and TTG (2 people), while a full 26% of CCP measurements were above the limit. No TPO measurements were above the limit. A convenient approach to incorporating such quantification limitations is by assuming a marginal proportional odds model and truncating the observations at the limit of quantification, which corresponds to (2.5) with ϵ_m coming from a logistic distribution. Finally, we used a Gaussian copula to estimate $\Sigma_{\epsilon\epsilon}$ and quantile-based bootstrap CIs to estimate variability in $\widehat{\boldsymbol{\theta}}$.

We estimated Σ_{hh} to be

$$\widehat{\Sigma}_{hh} = \begin{pmatrix} 1.25 & 0.20 & 0.21 & 0.22 \\ 0.20 & 1.48 & 0.25 & 0.04 \\ 0.21 & 0.25 & 1.45 & -0.13 \\ 0.22 & 0.04 & -0.13 & 1.20 \end{pmatrix},$$

suggesting that all autoantibodies were moderately positively correlated, with the exception of CCP and TPO, which exhibited a very small positive covariance, and TPO and TTG, which were moderately negatively correlated. The canonical correlations were estimated to be 0.63, 0.54, 0.39, and 0.38, with proportions of variance explained 0.091, 0.098, 0.033, and 0.033, for the first through fourth directions, respectively. Thus, there is a clear separation between the first two sCCA directions, which explain a much larger proportion of the

variance, and the last two directions.

The results for the first two sCCA directions are depicted in figure 2.1. In that figure, we see that the primary sCCA autoantibody direction is driven primarily by a positive association with TPO and a negative association with CCP. Contributing slightly less to the first direction is a positive association with ANA, while TTG contributes a small amount in the positive direction. The second direction is driven primarily by positive association with CCP, and secondarily by positive associations to the other three autoantibodies. Thus, we can potentially construe the first direction as non-CCP-related autoimmunity and the second direction as CCP-related autoimmunity. For the first direction, the 95% confidence intervals (CI) for ANA, CCP, and TPO all exclude 0, while for the second direction, only CCP's CI excludes 0. The third direction is driven primarily by a positive association with TTG, and the fourth direction is driven primarily by a positive association with ANA and a weaker negative association to TTG, though all CIs for these directions include 0.

The risk profile for non-CCP-related autoimmunity is driven primarily by rs2187668, which had previously shown associations to SLE (Taylor et al., 2011) and celiac (van Heel et al., 2007) and is located in the MHC region which is known to affect immune function. The 95% CI for the contribution of rs2187668 also excludes 0. Other SNPs who have CIs that exclude 0 are rs2366293 and rs10742326, both previously associated with SLE (Graham et al., 2011; Gateva et al., 2009). Female gender also had a CI that excluded 0.

On the other hand, the risk profile for CCP-related autoimmunity is driven primarily by rs1861525 on the CYCS gene which also had been previously linked to SLE (Gateva et al., 2009). Other important SNPs include rs3129860 (MHC region, previously associated with SLE (Taylor et al., 2011)), rs6457620 (previously associated with multiple sclerosis (Hafler et al., 2007) and RA (Denny et al., 2010)), rs6679677 (on RSBN1, previously associated with RA (Burton et al., 2007) and hyperthyroidism (Eriksson et al., 2012)), and rs3087243 (CTLA4 gene, previously associated with both autoimmune thyroid disease (Ueda et al., 2003) and RA (Plenge et al., 2005)). All the CIs for the listed SNPs excluded 0.

We finally present plots of the ranks of each y_m smoothed against the first two risk

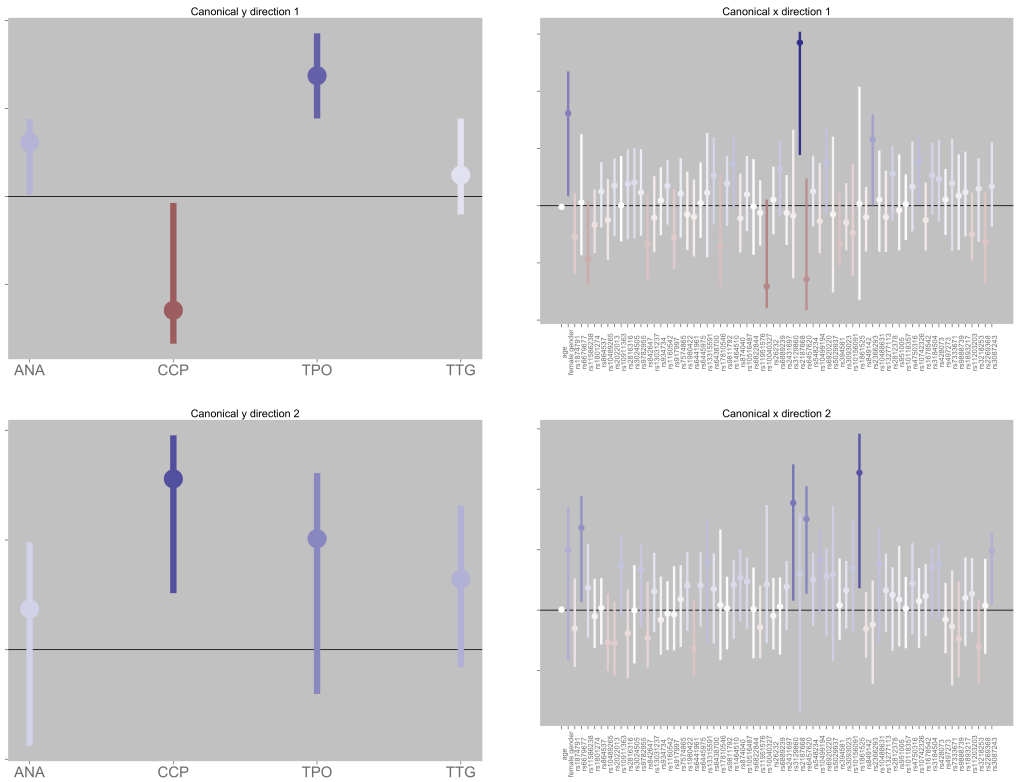


Figure 2.1: First two canonical directions for $\mathbf{x}$ and $\mathbf{y}$ in the autoantibody study. Height and color of points represent the contribution of each component to the canonical direction. Black line indicates 0. Line ranges represent bootstrap estimates of 95% confidence intervals. Scale is not listed because the directions are identifiable only up to a scaling constant.

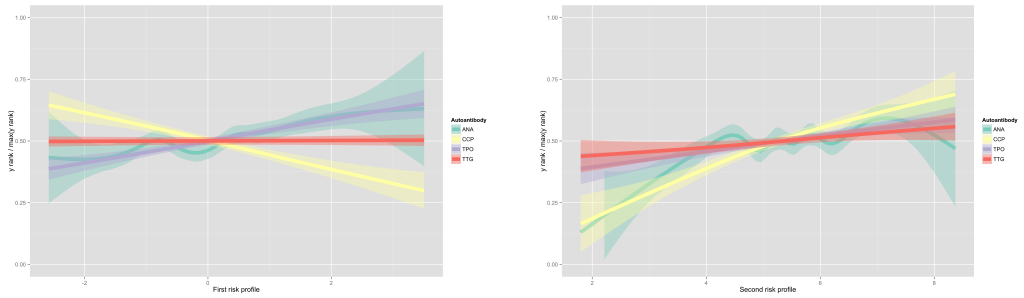


Figure 2.2: Nonparametric smoothing of relative rank of autoantibody measure against risk profile. Color indicates autoantibody: ANA (green), CCP (yellow), TTG (gray), TPO(red). The first risk profile is in the top left, the second is top right, and the third is bottom left.

profiles to examine the predictive value of the risk profiles apart from the chosen Gaussian copula. The use of ranks makes sense because those with very high and very low values of the autoantibodies do not have interpretable values apart from their rank. Ranks are normalized as $\text{rank}(y_{im}) \,/\, \max_i \text{rank}(y_{im})$ so that scales are equivalent across all autoantibodies. In figure 2.2 the smoothed plots are depicted, and we recover associations similar to those that would be expected by examining the canonical directions. For example, we saw from figure 2.1 that the first canonical direction was driven primarily by a strong positive association to TTG and a strong negative association to CCP. Figure 2.2 confirms this – those with high values of the risk profile tend to have lower ranks of CCP and higher ranks of TTG. Similarly, the second risk profile shows a strong positive association with CCP and weaker positive associations with the other three autoantibodies. Thus, we see that deriving the risk profiles from the Gaussian copula gives a reasonable result that is consistent with the underlying data.

2.3.2 Simulation results

We performed simulations to investigate the finite sample performance of sCCA. We generated data under the Gaussian copula model, where, for a given $\Sigma_{\epsilon\epsilon}$ and Σ_{xx}, we generated $\boldsymbol{\epsilon} \sim \Phi_{\Sigma_{\epsilon\epsilon}}$ and $\mathbf{x} \sim \Phi_{\Sigma_{xx}}$ and

$$z_m = h_m^*(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_{m0} + \epsilon_m).$$

We used $M = 4$ outcomes and $p = 10$ predictors. The transformations were chosen to be $h_1^*(x) = e^x, h_2^*(x) = x^3, h_3^*(x) = x$, and $h_4^*(x) = -\exp\{-e^x\}$. To demonstrate sCCA's ability to handle diverse $\mathbf{y}$, we thresholded y_2 and y_3 at 0, creating binary variables:

$$y_m = \begin{cases} z_m, & m = 1, 4 \\ 1, & z_m > 0, m = 2, 3 \\ 0, & z_m \leq 0, m = 2, 3. \end{cases}$$

We considered sample sizes of $n = 100$ and $n = 500$, and we considered various covariance structures for $\Sigma_{\epsilon\epsilon}$, including independence ($\Sigma_{\epsilon\epsilon} = I$), exchangeable ($\Sigma_{\epsilon\epsilon} = 0.3\mathbf{1}\mathbf{1}^\intercal + 0.7I$), and unstructured

$$\Sigma_{\epsilon\epsilon} = \begin{bmatrix} 1 & 0.16 & -0.12 & 0.49 \\ 0.15 & 1 & -0.01 & 0.36 \\ -0.12 & -0.01 & 1 & 0.02 \\ 0.49 & 0.36 & 0.02 & 1 \end{bmatrix}.$$

For ease of presenting simulation results, we will refer to the unique off-diagonal elements of $\Sigma_{\epsilon\epsilon}$ as $\boldsymbol{\sigma}_\epsilon = (\sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{23}, \sigma_{24}, \sigma_{34})^\intercal$, which for the unstructured case would be $(0.15, -0.12, 0.49, -0.01, 0.36, 0.02)^\intercal$.

Similarly, for Σ_{xx}, we considered two different structures: exchangeable ($\Sigma_{xx} = 0.1 \times (0.3\mathbf{1}\mathbf{1}^\intercal + 0.7I)$) and unstructured

$$\Sigma_{xx} = 0.1\times \begin{bmatrix} 1.00 & -0.06 & -0.13 & 0.09 & -0.18 & -0.42 & 0.51 & -0.08 & -0.02 & -0.01 \\ -0.06 & 1.00 & -0.05 & -0.01 & -0.03 & -0.07 & -0.03 & -0.02 & -0.03 & 0.00 \\ -0.13 & -0.05 & 1.00 & -0.02 & -0.10 & -0.24 & -0.05 & -0.05 & -0.08 & -0.01 \\ 0.09 & -0.01 & -0.02 & 1.00 & -0.02 & -0.05 & 0.07 & -0.01 & 0.00 & 0.00 \\ -0.18 & -0.03 & -0.10 & -0.02 & 1.00 & -0.14 & -0.10 & -0.03 & -0.06 & -0.01 \\ -0.42 & -0.07 & -0.24 & -0.05 & -0.14 & 1.00 & -0.25 & -0.07 & -0.13 & -0.02 \\ 0.51 & -0.03 & -0.05 & 0.07 & -0.10 & -0.25 & 1.00 & -0.05 & 0.00 & 0.00 \\ -0.08 & -0.02 & -0.05 & -0.01 & -0.03 & -0.07 & -0.05 & 1.00 & -0.03 & 0.00 \\ -0.02 & -0.03 & -0.08 & 0.00 & -0.06 & -0.13 & 0.00 & -0.03 & 1.00 & -0.01 \\ -0.01 & 0.00 & -0.01 & 0.00 & -0.01 & -0.02 & 0.00 & 0.00 & -0.01 & 1.00 \end{bmatrix}.$$

To demonstrate the performance of the bootstrap, we performed 1000 bootstrap samples for each simulation setting for which Σ_{xx} is exchangeable.

We first examine the bias in our estimators. In figure 2.3, we depict the absolute bias of our estimators. For ease of presentation, we consider the median absolute deviation of $\widehat{\mathbf{b}}_k$ and $\widehat{\boldsymbol{a}}_k$: $m_{\widehat{b}_k} = \text{median}\Big\{(|\widehat{b}_{kj} - b_{kj})|)_{1\leq j\leq 4}\Big\}$. We see from the plot that, as expected, the biases tend to be higher for smaller sample sizes and for later directions, e.g., the biases tend

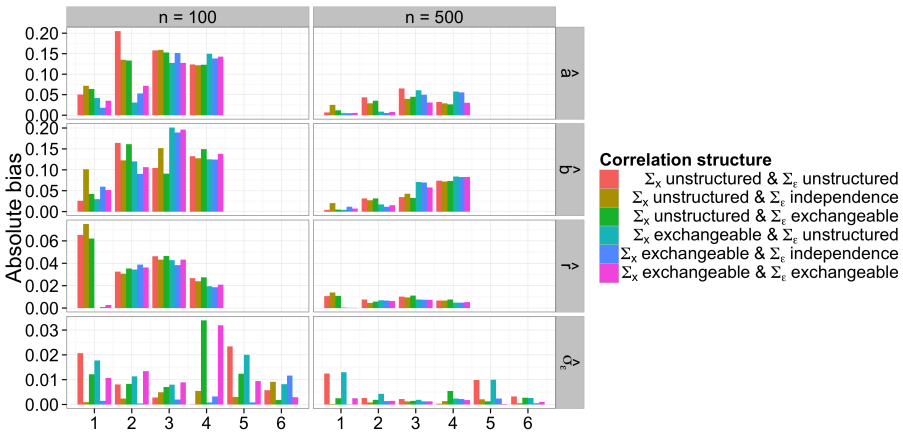


Figure 2.3: Absolute bias of sCCA estimators, represented by the height of bars. Color denotes simulation setting, i.e. correlation structures of $\mathbf{x}$ and $\boldsymbol{\epsilon}$. Panels in the left column represent sample size of 100, while panels on the right represent $n = 500$. Results for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k, \widehat{r}_k$, and σ_ϵ are depicted in the first through fourth rows, respectively. The x-axis denotes denotes canonical direction (k) for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k$, and $\widehat{r}_k$, while the x-axis denotes the entry in the vector for σ_ϵ. Median absolute bias is given for $\widehat{\mathbf{b}}_k$ and $\widehat{\boldsymbol{a}}_k$.

to be higher for direction 3 than for directions 2 and 1. The bias in the estimation of σ_ϵ is small regardless of correlation structure and is nearly nonexistent when $n = 500$. In general, it appears that there is more bias in estimation of first direction parameters when Σ_{xx} is unstructured rather than exchangeable.

Much of the bias depicted in 2.3 can be attenuated by adjusting each $\widehat{\theta}_j$ by the estimated bias $\widehat{\text{bias}}_{\widehat{\theta}_j}$

$$\widehat{\theta}_{j,\text{adj}} = \widehat{\theta}_j - \widehat{\text{bias}}_{\widehat{\theta}_j}.$$

Figure 2.4 depicts absolute bias for the bias-adjusted sCCA estimators $\widehat{\theta}_{j,\text{adj}}$ and shows a dramatic reduction in the magnitude of bias. However, the estimates of the third and fourth canonical directions may still incur relatively large amounts of bias in practice.

We also examined the performance of the bootstrap in estimating the variability in $\widehat{\boldsymbol{\theta}}$. In figure 2.5, we depict the performance of bootstrap standard errors, and in figure 2.6 we depict the performance of bootstrap confidence intervals. For ease of presentation, the median coverage is given for the components of $\widehat{\boldsymbol{a}}_k$ and $\widehat{\mathbf{b}}_k$. We can see that standard error estimates are in general quite accurate for the empirical standard errors, with most

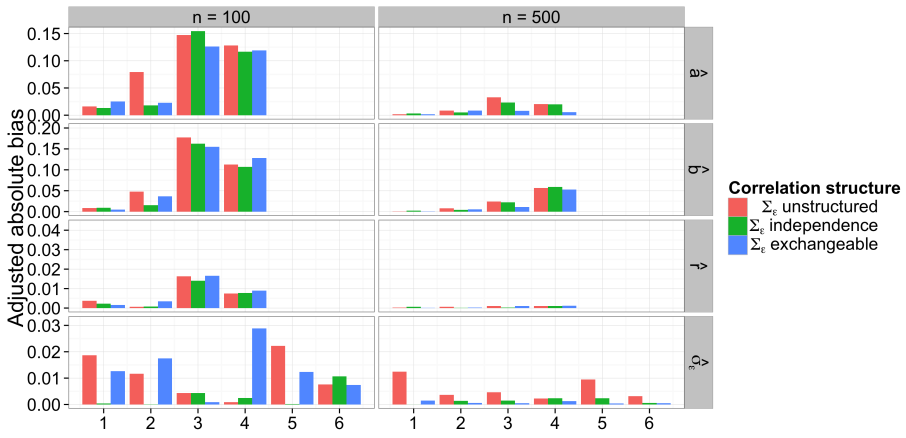


Figure 2.4: Adjusted absolute bias of sCCA estimators, represented by the height of bars. Color denotes simulation setting, i.e. correlation structure of $\boldsymbol{\epsilon}$ (the correlation structure of $\mathbf{x}$ is exchangeable). Panels in the left column represent sample size of 100, while panels on the right represent $n = 500$. Results for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k, \widehat{r}_k$, and $\widehat{\sigma}_\epsilon$ are depicted in the first through fourth rows, respectively. The x-axis denotes denotes canonical direction (k) for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k$, and $\widehat{r}_k$, while the x-axis denotes the entry in the vector for σ_ϵ. Median absolute bias is given for $\widehat{\mathbf{b}}_k$ and $\widehat{\boldsymbol{a}}_k$.

dots falling on or near the dotted line that represents perfect estimation. Importantly, for direction 1, bootstrap estimates of the standard error, the only significant departures from the dotted line correspond to overestimates – for $\widehat{\boldsymbol{a}}_1$ when $n = 100$. Thus, if anything, the bootstrap is conservative in estimating standard errors. It seems that for later directions, the bootstrap may at times underestimate the standard error, especially for $\widehat{\mathbf{b}}_k$.

We use percentile confidence intervals for all measures besides r_k, for which we use the basic interval. When $n = 500$, confidence interval coverage is quite good for all estimators except for $\widehat{\boldsymbol{a}}_4$, and $\widehat{\mathbf{b}}_4$, which indicates once again that performance of sCCA estimates may degrade once more than a few directions are obtained. Confidence interval coverage is similar for $n = 100$, with disappointing coverage for $\widehat{\boldsymbol{a}}_1$ and $\widehat{r}_2$.

To get a more granular picture of the performance of sCCA, we depict results in detail for Σ_{xx} exchangeable and $\Sigma_{\epsilon\epsilon}$ unstructured. In figure 2.7 we illustrate results for the canonical directions with $n = 500$. Plots depict averages of $\widehat{\boldsymbol{a}}_k$ and $\widehat{\mathbf{b}}_k$ across the 1000 simulations. Red dots depict the average estimates, and blue dots depict average bias-corrected estimates. Horizontal bars depict true parameter values. Black vertical bars depict the em-

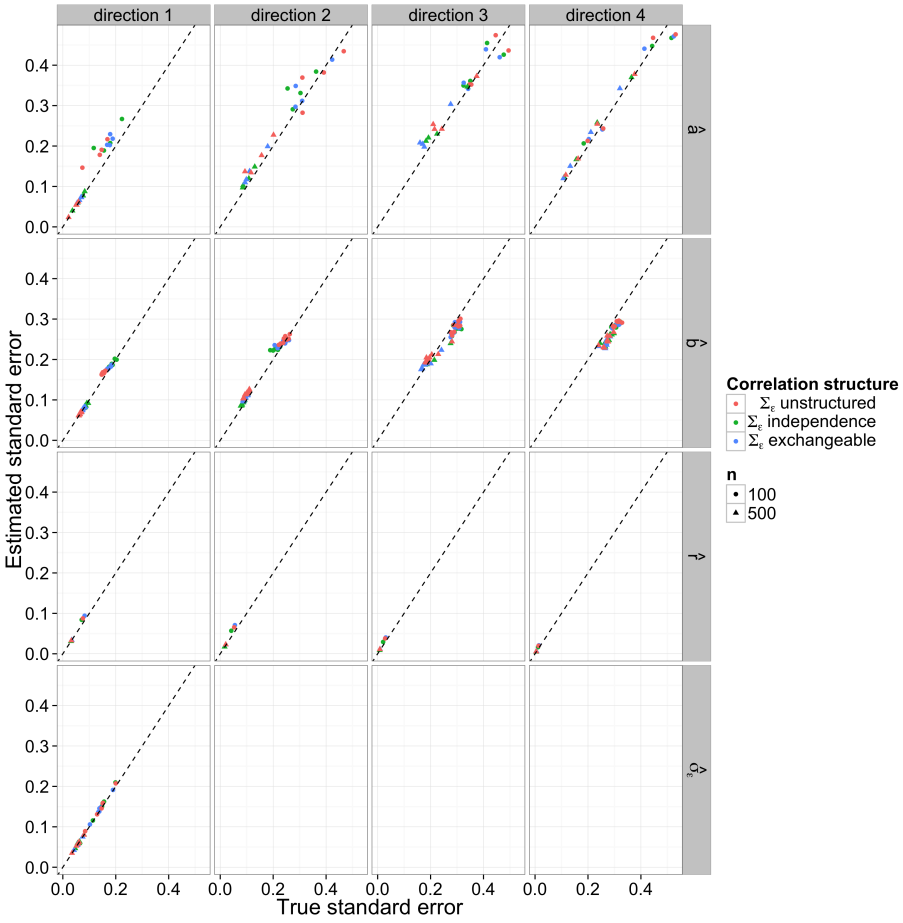


Figure 2.5: Estimated bootstrap standard errors versus empirical standard errors. Color denotes simulation setting, i.e. correlation structure of $\boldsymbol{\epsilon}$ (the correlation of $\mathbf{x}$ is exchangeable). Columns represent canonical direction, all components of $\boldsymbol{\sigma}_\epsilon$ are depicted under direction 1. Results for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k, \widehat{r}_k$, and $\widehat{\sigma}_\epsilon$ are depicted in the first through fourth rows, respectively.

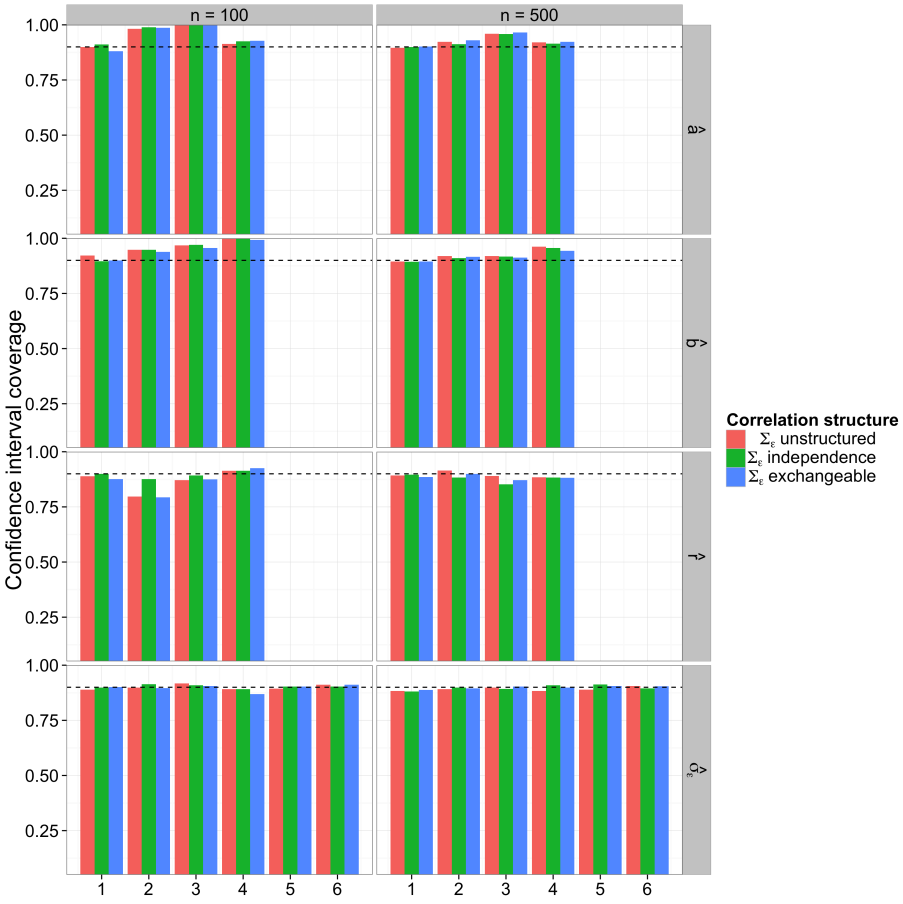


Figure 2.6: Confidence interval coverage for sCCA estimators. Color denotes simulation setting, i.e. correlation structure $\boldsymbol{\epsilon}$ (the correlation structure of $\mathbf{x}$ is exchangeable). Panels in the left column represent sample size of 100, while panels on the right represent $n = 500$. Results for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k, \widehat{r}_k$, and $\widehat{\sigma}_\epsilon$ are depicted in the first through fourth rows, respectively. The x-axis denotes denotes canonical direction (k) for $\widehat{\boldsymbol{a}}_k, \widehat{\mathbf{b}}_k$, and $\widehat{r}_k$, while the x-axis denotes the entry in the vector for $\widehat{\sigma}_\epsilon$. Median coverage is given for $\widehat{\boldsymbol{a}}_k$ and $\widehat{\mathbf{b}}_k$.

pirical 5%/95% quantiles of the estimates, which we will refer to as empirical intervals, and blue vertical bars depict average bootstrap confidence intervals.

We can see that the point estimates compare well to the true parameters for the first two directions. Performance of the sCCA estimates degrades as the third and fourth directions are obtained, both in terms of bias and size of empirical intervals. However, bias-corrected estimates remain close to the true parameter values even in the later directions.

Furthermore, we illustrate results for $\widehat{\Sigma}_{\epsilon\epsilon}$ and $\widehat{r}_k$ in figure 2.8. We see that the estimates of $\widehat{r}_k$ and the components of $\widehat{\Sigma}_{\epsilon\epsilon}$ compare quite favorably to their true values.

At smaller sample sizes, performance degrades much faster. In figure 2.9, we see that the first direction is once again quite accurate, but the other three directions are much less accurate, with similar results for $\widehat{\mathbf{b}}_k$. Estimation of $\Sigma_{\epsilon\epsilon}$ remains quite good at the lower sample size, but estimation of r_k is also worse after the first direction.

2.4 Discussion

We have proposed semiparametric canonical correlation analysis, a method to relate a set of covariates $\mathbf{x}$ to a diverse $\mathbf{y}$. That is, sCCA allows researchers to build risk profiles for many related phenotypes, even when those phenotypes are on incomparable scales, measured in completely different ways, or incompletely observed. Moreover, the canonical directions may offer insight into clustering or relatedness in $\mathbf{x}$ and $\mathbf{y}$ or into subgroups within the data.

The assumption of 2.5 should not be considered restrictive, since, for a given g_m in the absence of the linear predictor $-\mathbf{x}^\intercal\boldsymbol{\beta}_{m0}$, some h_{m0} can always be found to satisfy the model. However, the necessity of the parametric copula function may raise worries of misspecification. To that end, we also propose nonparametric smoothing as a way to verify the predictive accuracy of the sCCA risk profiles in the event of misspecification of the copula function.

We have demonstrated sCCA's potential by applying it to a study to identify risk profiles for autoimmune disease. We identified risk profiles for two potential types of autoimmune traits – non-CCP autoimmunity and CCP-related autoimmunity – and we verified them using the nonparametric smoothing method.

Our sCCA methods as proposed above are not suitable to high-dimensional $\mathbf{x}$, since

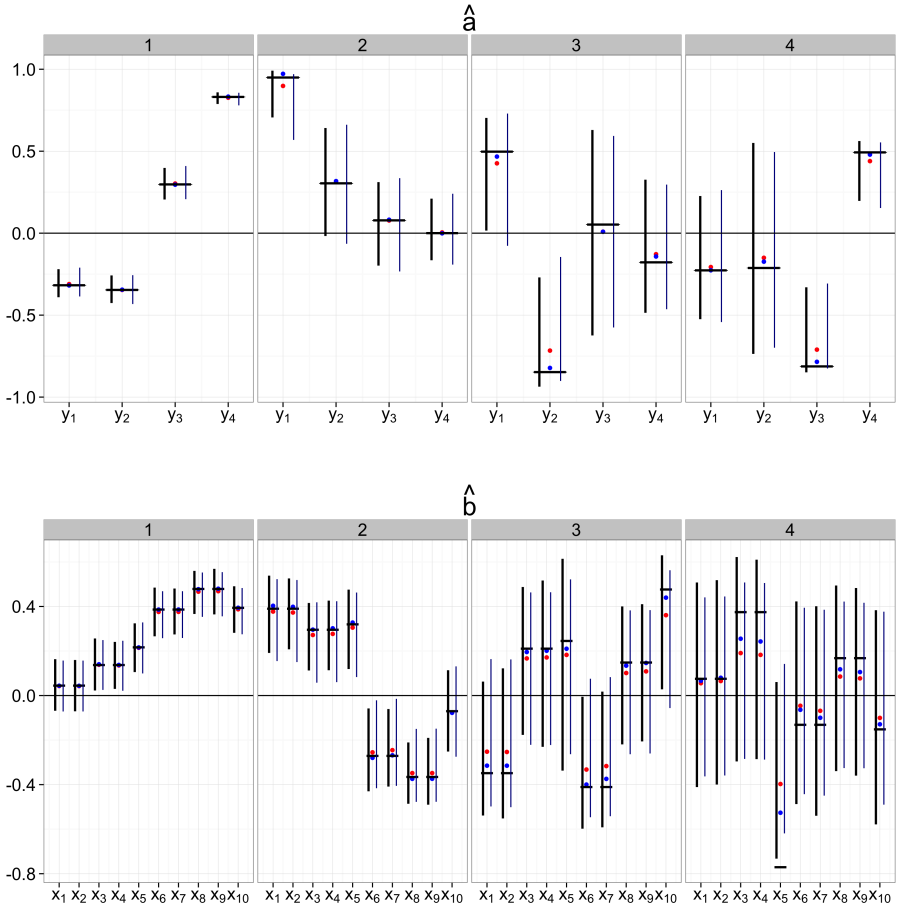


Figure 2.7: Average parameter estimates and empirical intervals for sCCA directions $\boldsymbol{a}$ and $\mathbf{b}$, with $n = 500, \Sigma_{xx}$ exchangeable, and $\Sigma_{\epsilon\epsilon}$ unstructured. All averages are taken over all 1000 simulations. Black dots represent average point estimates, and black lines represent empirical 90% intervals. Horizontal lines represent true parameter values. The first panel represents the first direction, the second panel represents the second direction, and so on.

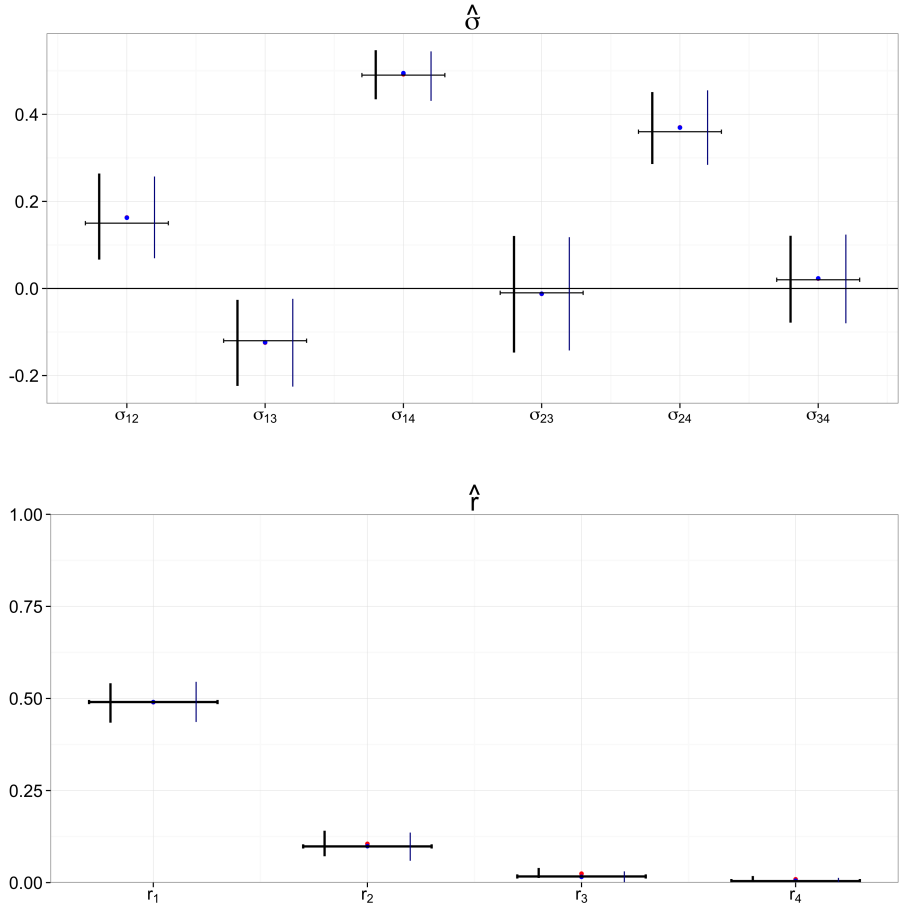


Figure 2.8: Average parameter estimates and empirical intervals for the proportion of variance explained r_k and the off-diagonal elements of $\Sigma_{\epsilon\epsilon}$, with $n = 500, \Sigma_{xx}$ exchangeable, and $\Sigma_{\epsilon\epsilon}$ unstructured. All averages are taken over all 1000 simulations. Black dots represent average point estimates, and black lines represent empirical 90% intervals. Horizontal lines represent true parameter values.

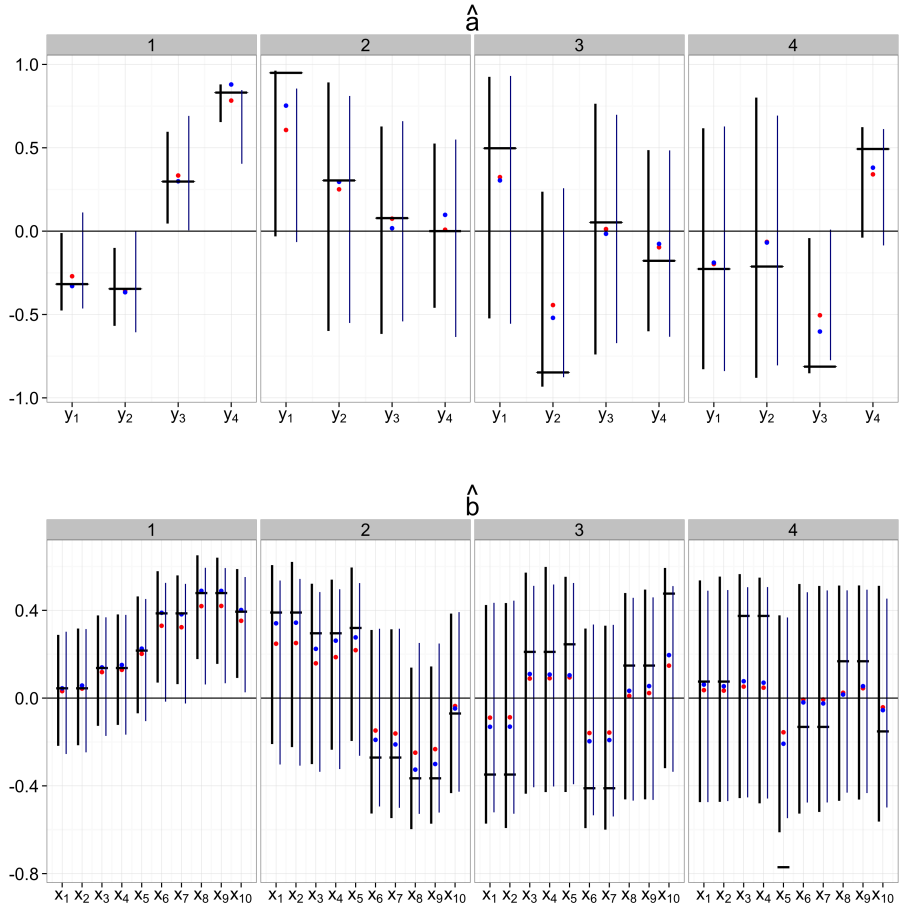


Figure 2.9: Average parameter estimates and confidence intervals for sCCA directions $\boldsymbol{a}$ and $\mathbf{b}$, with $n = 100, \Sigma_{xx}$ exchangeable, and $\Sigma_{\epsilon\epsilon}$ unstructured. All averages are taken over all 1000 simulations. Black dots represent average point estimates, and black lines represent empirical 90% intervals. Horizontal lines represent true parameter values. The first panel represents the first direction, the second panel represents the second direction, and so on.

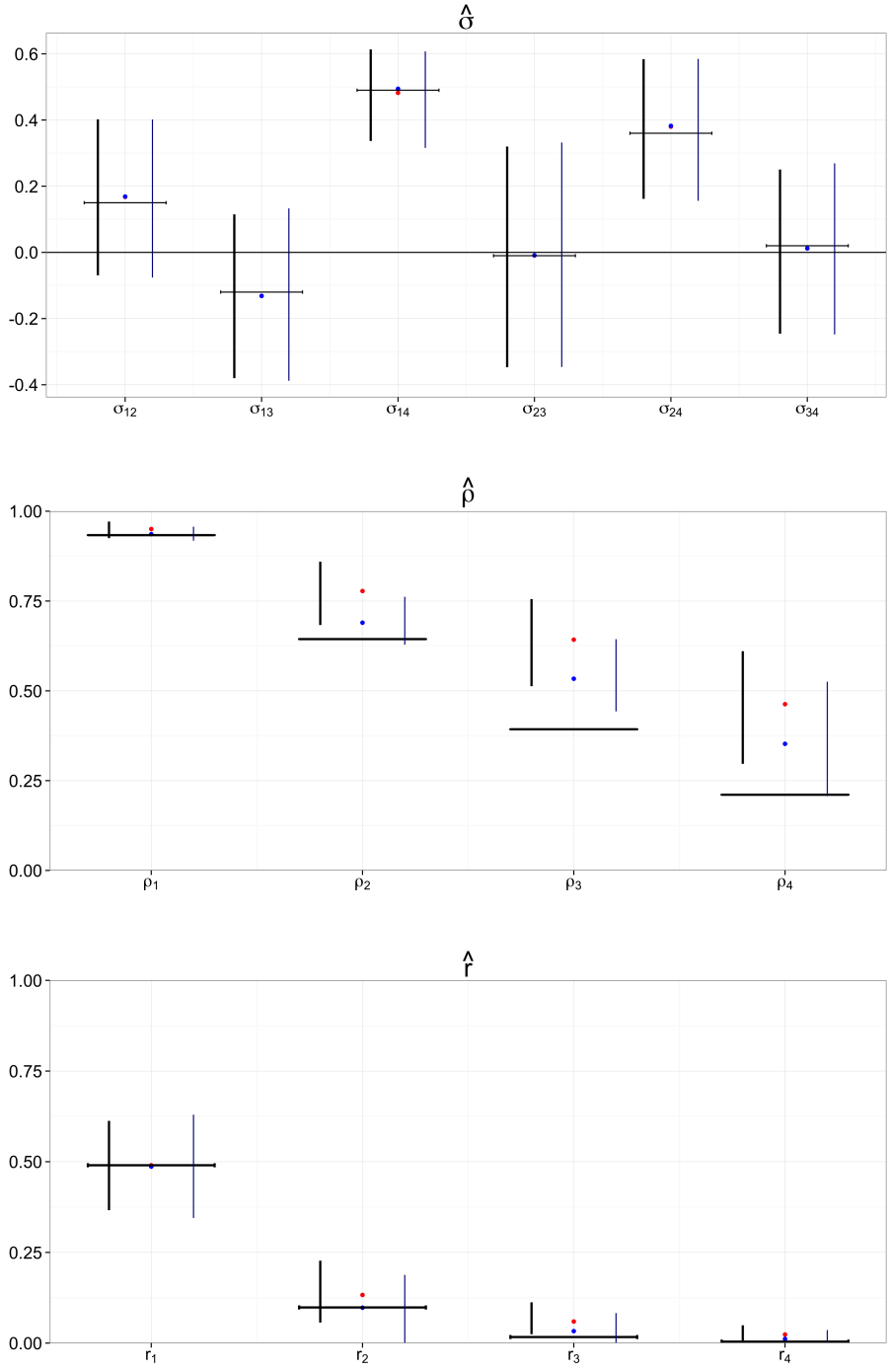


Figure 2.10: Average parameter estimates and confidence intervals for the proportion of variance explained r_k and the off-diagonal elements of $\widehat{\Sigma}_{\epsilon\epsilon}$, with $n = 100, \Sigma_{xx}$ exchangeable, and $\Sigma_{\epsilon\epsilon}$ unstructured. All averages are taken over all 1000 simulations. Black dots represent average point estimates, and black lines represent empirical 90% intervals. Horizontal lines represent true parameter values.

the performance of the model 2.5 and standard covariance matrix estimates degrade as the dimension of $\mathbf{x}$ grows close to the sample size n. However, there are many interesting clinical situations where considering high dimensional $\mathbf{x}$ may be desirable, most notably in genomics where the number of potential markers can be quite large. In these cases, one may want to employ sparse estimators to improve sCCA, similar to what has been done in Witten and Tibshirani (2009) or Zou et al. (2006).

2.5 Appendix

2.5.1 Expansions of $\widehat{\boldsymbol{\sigma}}$

In this section, we show that it is possible to write $n^{-\frac{1}{2}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$ as a sum of iid terms, which ensures that all of the statistics of interest are asymptotically normal. There are three types of components of $\widehat{\boldsymbol{\sigma}}$, those coming from $\widehat{\Sigma}_{xx}$, those coming from $\widehat{\Sigma}_{hx}$, and those coming from $\widehat{\Sigma}_{hh}$. We will consider them each in turn.

The components of $n^{\frac{1}{2}}(\widehat{\Sigma}_{xx} - \Sigma_{xx})$ can be written

$$\begin{aligned}
&n^{\frac{1}{2}}(n-1)^{-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_{ik}) - n^{\frac{1}{2}}\sigma_{xxjk} \\
&\quad = n^{-\frac{1}{2}}\sum_{i=1}^{n}(x_{ij} - \mu_j)(x_{ik} - \mu_k) - n^{\frac{1}{2}}\sigma_{xxjk} - n^{\frac{1}{2}}(\bar{x}_j - \mu_j)(\bar{x}_k - \mu_k) + o_p(1) \\
&\quad = n^{\frac{1}{2}}\left\{n^{-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)(x_{ik} - \mu_k) - \sigma_{xxjk}\right\} + o_p(1)
\end{aligned}$$

for $j, k \in \{1, ..., p\}$ and $\sigma_{xxjk} = (\Sigma_{xx})_{jk}$.

The components of $n^{\frac{1}{2}}(\widehat{\Sigma}_{hx} - \Sigma_{hx})$ can be written

$$\begin{aligned}
n^{\frac{1}{2}}(\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} - \Sigma_{xx}\mathbb{B}_0) &= n^{\frac{1}{2}}(\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} - \Sigma_{xx}\widehat{\mathbb{B}}) + n^{\frac{1}{2}}(\Sigma_{xx}\widehat{\mathbb{B}} - \Sigma_{xx}\mathbb{B}_0) \\
&= n^{\frac{1}{2}}\left\{n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\intercal} - \boldsymbol{\sigma}_{xx}\right\}\widehat{\mathbb{B}} + \Sigma_{xx}\left\{n^{\frac{1}{2}}(\widehat{\mathbb{B}} - \mathbb{B}_0)\right\} \\
&= n^{\frac{1}{2}}\left\{n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\intercal} - \boldsymbol{\sigma}_{xx}\right\}\mathbb{B}_0 + \Sigma_{xx}\left\{n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathcal{U}_i(\mathbb{B}_0)\right\} + o_p(1) \\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^{\intercal}\mathbb{B}_0 + \Sigma_{xx}\mathcal{U}_{\beta i}(\mathbb{B}_0) - \sigma_{xx}\right\} + o_p(1)
\end{aligned}$$

where $\mathcal{U}_{\beta i}(\mathbb{B}_0)$ is a $p \times K$ matrix with columns $\mathcal{U}_{\beta i}(\boldsymbol{\beta}_{m0})$ and $n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{U}_{\beta i}(\boldsymbol{\beta}_{m0})$ is the influence function corresponding to $\widehat{\boldsymbol{\beta}}_m$ from model (2.5).

And the components of $n^{\frac{1}{2}}(\widehat{\Sigma}_{hh} - \Sigma_{hh})$ can be written

$$n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} + \widehat{\Sigma}_{\epsilon\epsilon} - \mathbb{B}_0^{\intercal}\Sigma_{xx}\mathbb{B}_0 + \widehat{\Sigma}_{\epsilon\epsilon}) = n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} - \mathbb{B}_0^{\intercal}\Sigma_{xx}\mathbb{B}_0) + n^{\frac{1}{2}}(\widehat{\Sigma}_{\epsilon\epsilon} - \Sigma_{\epsilon\epsilon})$$

The elaboration of the first term follows a similar form as the elaboration of $n^{\frac{1}{2}}(\widehat{\Sigma}_{hx} - \Sigma_{hx})$,

$$\begin{aligned}
n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} - \mathbb{B}_0^{\intercal}\Sigma_{xx}\mathbb{B}_0) &= n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\widehat{\Sigma}_{xx}\widehat{\mathbb{B}} - \widehat{\mathbb{B}}^{\intercal}\Sigma_{xx}\widehat{\mathbb{B}}) + n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\Sigma_{xx}\widehat{\mathbb{B}} - \widehat{\mathbb{B}}^{\intercal}\Sigma_{xx}\mathbb{B}_0) + \\
&\quad n^{\frac{1}{2}}(\widehat{\mathbb{B}}^{\intercal}\Sigma_{xx}\mathbb{B}_0 - \mathbb{B}_0^{\intercal}\Sigma_{xx}\mathbb{B}_0) \\
&= \widehat{\mathbb{B}}^{\intercal}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\{(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^{\intercal}\mathbb{B}_0 - \sigma_{xx}\}\right)\widehat{\mathbb{B}} \\
&\quad + \widehat{\mathbb{B}}^{\intercal}\Sigma_{xx}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathcal{U}_{\beta i}(\mathbb{B}_0)\right) + \left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathcal{U}_{\beta i}(\mathbb{B}_0)\right)^{\intercal}\Sigma_{xx}\mathbb{B}_0 + o_p(1) \\
&= \mathbb{B}_0^{\intercal}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\{(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^{\intercal}\mathbb{B}_0 - \sigma_{xx}\}\right)\mathbb{B}_0 + \\
&\quad 2\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathcal{U}_{\beta i}(\mathbb{B}_0)\right)^{\intercal}\Sigma_{xx}\mathbb{B}_0 + o_p(1) \\
&= n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{\mathbb{B}_0^{\intercal}\left((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^{\intercal}\mathbb{B}_0 - \sigma_{xx}\right)\mathbb{B}_0 + 2\left(\mathcal{U}_{\beta i}(\mathbb{B}_0)\right)^{\intercal}\Sigma_{xx}\mathbb{B}_0\right\} \\
&\quad + o_p(1)
\end{aligned}$$

For the second term, consider the function that is solved to obtain $\widehat{\Sigma}_{\epsilon\epsilon}$. Let $\widehat{\boldsymbol{\Psi}}_n(\Sigma) = n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \widehat{\mathcal{H}}(\mathbf{y}_i), \widehat{\mathbb{B}})$ be the score equations based on $\mathcal{L}(\Sigma_{\epsilon\epsilon}; \mathbb{V}, \widehat{\mathcal{H}}, \widehat{\mathbb{B}})$ used to obtain $\widehat{\Sigma}_{\epsilon\epsilon}$. The notation $\widehat{\mathcal{H}}(\mathbf{y}_i)$ indicates that the value at which $\widehat{h}_m(\cdot)$ is evaluated depends on y_{im} (the dependence on $\boldsymbol{\delta}_i$ is suppressed).

$$\begin{aligned}
\widehat{\boldsymbol{\Psi}}_n(\Sigma) &= n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \widehat{\mathcal{H}}(\mathbf{y}_i), \widehat{\mathbb{B}}) \\
&= n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \mathcal{H}_0(\mathbf{y}_i), \mathbb{B}_0) + n^{-1}\sum_{i=1}^{n}\dot{\boldsymbol{\psi}}(\Sigma; \mathbb{V}_i, \mathcal{H}_0(\mathbf{y}_i), \mathbb{B}_0)^{\intercal}(\widehat{\boldsymbol{\gamma}}(\mathbf{y}_i) - \boldsymbol{\gamma}_0(\mathbf{y}_i)) + \\
&\quad \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2
\end{aligned}$$

where $\dot{\boldsymbol{\psi}}(\Sigma; \mathbb{V}_i, \mathcal{H}, \mathbb{B})$ is the derivative of $\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \mathcal{H}, \mathbb{B})$ with respect to $\boldsymbol{\gamma} = (\beta_{jm}, h_m(y_{im}))_{1\leq m\leq M, 1\leq j\leq p}$, $\gamma_0(\mathbf{y}_i) = (\beta_{jm0}, h_{m0}(y_{im}))_{1\leq m\leq M, 1\leq j\leq p}$, and $\widehat{\boldsymbol{\gamma}}(\mathbf{y}_i) = (\widehat{\beta}_{jm}, \widehat{h}_m(y_{im}))_{1\leq m\leq M, 1\leq j\leq p}$. Now, let $\mathbb{V}(t) = (\mathbf{t}_y, \mathbf{t}_\delta, \mathbf{t}_x)$, then

$$\begin{aligned}\widehat{\boldsymbol{\Psi}}_n(\Sigma) = n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \mathcal{H}_0(\mathbf{y}_i), \mathbb{B}_0) + \\ \int \dot{\boldsymbol{\psi}}(\Sigma; \mathbb{V}(t), \mathcal{H}_0(\mathbf{t}_y), \mathbb{B}_0)^\intercal(\widehat{\boldsymbol{\gamma}}(\mathbf{t}_y) - \boldsymbol{\gamma}_0(\mathbf{t}_y))d\left\{n^{-1}\sum_{i=1}^{n} I_{\mathbb{V}(\mathbf{t})\leq\mathbb{V}_i}\right\} \\ + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2\end{aligned}$$

Because $n^{-1}\sum_{i=1}^{n} I_{\mathbb{V}(\mathbf{t})\leq\mathbb{V}_i}$ has bounded total variation we have from (Zeng and Lin, 2010) that

$$\begin{aligned}\widehat{\Psi}_n(\Sigma) &= n^{-1}\sum_{i=1}^{n}\{\boldsymbol{\psi}(\Sigma; \mathbb{V}_i, \mathcal{H}_0(\mathbf{y}_i), \mathbb{B}_0) + \mathcal{U}_{\gamma i}(\gamma_0)\} + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2 \\ &= n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}_i^*(\Sigma) + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2 = \boldsymbol{\Psi}_n^*(\Sigma) + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2\end{aligned}$$

where $\mathcal{U}_{\gamma i}(\gamma_0)$ is a mean-zero function of the influence function for $\widehat{\gamma}$. This means that, the components of $n^{\frac{1}{2}}(\widehat{\Sigma}_{\epsilon\epsilon} - \Sigma_{\epsilon\epsilon})$ can be written as

$$-n^{-\frac{1}{2}}\sum_{i=1}^{n}\dot{\boldsymbol{\Psi}}^*{}_n(\Sigma_{\epsilon\epsilon})^{-1}\boldsymbol{\psi}_i^*(\Sigma_{\epsilon\epsilon}) + o_p(1).$$

So, finally, the components of $n^{\frac{1}{2}}(\widehat{\Sigma}_{hh} - \Sigma_{hh})$ can be written

$$\begin{aligned}n^{-\frac{1}{2}}\sum_{i=1}^{n}\Big[\mathbb{B}_0^\intercal\left((\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^\intercal\mathbb{B}_0 - \sigma_{xx}\right)\mathbb{B}_0 + 2\left(\mathcal{U}_{\beta i}(\mathbb{B}_0)\right)^\intercal\Sigma_{xx}\mathbb{B}_0 - \dot{\boldsymbol{\Psi}}^*{}_n(\Sigma_{\epsilon\epsilon})^{-1}\boldsymbol{\psi}_i^*(\Sigma_{\epsilon\epsilon})\Big] \\ + o_p(1)\end{aligned}$$

And the demonstration that $n^{-\frac{1}{2}}(\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma})$ can be written as a sum of iid terms is concluded.

Genome-wide association studies of longitudinal outcomes

Denis Agniel and Tianxi Cai

Department of Biostatistics

Harvard School of Public Health

3.1 Introduction

An important goal of genome-wide association studies (GWAS) is to explore susceptibility to complex diseases. These studies have led to identification of many genomic regions as putatively harboring disease susceptibility alleles for a wide range of disorders. For patients with a particular disease, GWAS have also been performed to identify genetic variants associated with progression of disease. The disease progression is often monitored by longitudinally measured biological markers. Such longitudinal measures allow researchers to more clearly characterize clinical outcomes that cannot necessarily be captured in one or even a few measurements. For example, trajectories of a measure called disease activity score (DAS) are often used to quantify the progression of rheumatoid arthritis (RA) (Prevoo et al., 1995), and forced expiratory volume can similarly be used to monitor the progression of lung health (Yang et al., 2009). For HIV-infected patients, the trajectories of HIV RNA (viral load) and $CD4^+$ T cell counts are important parameters for monitoring the disease and determining optimal treatment strategies.

The most common approach to analyzing longitudinal data of this kind is to use linear mixed effects (LME) models (Laird and Ware, 1982), which relate the longitudinal outcome $\mathbf{y} = (y_1, ..., y_r)^\intercal$, measured at times $\mathbf{t} = (t_1, ..., t_r)^\intercal$, linearly to genetic characteristics $\mathbf{z}$, covariates $\mathbf{x}$, and $\mathbf{t}$ with both fixed and random effects. However, in general, such a linear relationship is likely to be overly simplistic, especially for markers such as DAS that may oscillate over time. To incorporate nonlinear trajectories, nonlinear and nonparametric mixed effects models as well as functional regression methods have been proposed (Davidian and Giltinan, 2003; Lindstrom and Bates, 1990; Rice and Wu, 2001; Wu and Zhang, 2002; Gertheiss et al., 2013; Guo, 2002; Ramsay, 2006, e.g.). These methods typically do not specify the functional form of the trajectories *a priori* but rely on nonparametric smoothing or basis function expansions. Some of these methods also require additional assumptions such as dense and/or regular measurement times. For example, the functional smoothing random effects model (Chiou et al., 2003) regresses a trajectory $\mathbf{y}$ onto $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{t}$, but requires $\mathbf{y}$ to be densely and regularly observed, a restriction which is not commonly met in the longitudinal setting. Another drawback of standard nonparametric approaches is that

the effective degrees of freedom (DF) tends to be quite large, which would in turn lead to low power in identifying important genomic variants. To effectively capture the nonlinearity with potentially low DF, functional principal component analysis (FPCA) methods have been proposed in recent years (Castro et al., 1986; Rice and Silverman, 1991; Yao et al., 2005; Hall et al., 2006). FPCA methods estimate a few leading eigenfunctions that can be used to approximate the space of the true trajectory functions with minimal assumptions on the functional form. However, most existing FPCA methods focus primarily on estimation and regression problems. Although some of these methods could be used to derive testing procedures, such procedures would require performing functional regression for each of the millions of variants in a GWAS, which would be computationally infeasible in most settings. Furthermore, these regression methods require restrictive normality assumptions which may lead to invalid tests when the assumptions are violated.

In this paper, employing FPCA along with a variance component testing framework, we propose a *Functional Principal Variance Component (FPVC)* testing procedure that can capture the nonlinear trajectories without requiring a normality assumption or fitting individual functional regression models. Specifically, we conceive of $\mathbf{y}$ as a noisy realization of a smooth underlying function $Y(\cdot)$, and we employ FPCA to identify the major patterns of variation in $Y(\cdot)$ by estimating its underlying eigenfunctions. We approximate $Y(\cdot)$ as a linear combination of these eigenfunctions. For each patient, his/her $Y(\cdot)$ is then approximated by a weighted average of the estimated eigenfunctions, with weights corresponding to functional principal component *loadings* or *scores*. Since longitudinal data tends to be sampled at irregular time intervals, we use the best linear unbiased predictor (BLUP) to estimate the scores. The BLUP was also the basis for FPCA with sparse longitudinal data in the *principal analysis via conditional expectation* (PACE) method (Yao et al., 2005) under a normality assumption. Here, we use BLUP to motivate our testing procedure but do not require the normality for the validity of the FPVC test. The test statistic can be derived through the variance component testing framework and viewed as a summary measure of the overall covariance between the estimated subject-specific scores, which characterize the person's trajectory, and the genetic markers. Similar variance component tests have previously been

proposed for standard linear and logistic regressions with observed single outcomes (Wu et al., 2011).

The primary virtues of FPVC testing are threefold. First, we separate the procedure into two stages of distinct complexity in order to make it feasible on the genome-wide scale. In the first stage, we model $\mathbf{y}$ flexibly using FPCA and obtain a succinct summary of $\mathbf{y}$ for each patient, once and for all. In the second stage, we perform a rather simple model on the genome-wide scale. Thus, we segregate the computationally complex stage (which need occur only once) from the genome-wide stage (which could require the same computation on the order of millions of times). Second, the summary of $\mathbf{y}$ that we obtain from FPCA is the most succinct summary possible, as the eigenfunctions identified by FPCA are the functions that explain the most variability in $\mathbf{y}$. Third, our theoretical results suggest that the null distribution of the FPVC test statistic reduces to a simple mixture of χ^2 distributions. The variability due to estimating the eigenfunctions does not induce any additional noise at the first order under the null, which greatly simplifies the estimation of the null distribution.

The structure of the rest of the paper is as follows. In section 3.2, we describe FPCA and introduce FPVC testing and our main results. In section 3.3, we discuss simulation results. In section 3.4, we apply our proposed method to a study to detect the effects of genetic markers on the progression of low-density lipoprotein (LDL) cholesterol. In section 3.5 we discuss further implications of our procedure.

3.2 Functional principal variance component testing

3.2.1 The test statistic

In this section, we propose a testing procedure for assessing the association between a set of genetic markers and a longitudinally measured outcome $\mathbf{y}$, adjusting for covariates. Let the data for analysis consist of n independent random vectors $\mathbb{V} = \{\mathbf{V}_i = (\mathbf{y}_i^\intercal, \mathbf{t}_i^\intercal, \mathbf{z}_i^\intercal, \mathbf{x}_i^\intercal)^\intercal\}_{i=1}^n$, where $\mathbf{y}_i = (y_{i1}, ..., y_{ir_i})^\intercal$ is a vector of outcome measurements taken at times $\mathbf{t}_i = (t_{i1}, ..., t_{ir_i})^\intercal \in \mathcal{T}^{r_i}$, $\mathcal{T}$ is a closed and bounded interval, $\mathbf{z}_i = (z_{i1}, ..., z_{ip})^\intercal$ is a vector of genetic markers of interest, and $\mathbf{x}_i = (1, x_{i1}, ..., x_{iq})^\intercal$ is a vector of additional covariates that are potentially related to the outcome, all measured on person i. For each i, we take $(\mathbf{z}_i^\intercal, \mathbf{x}_i^\intercal)^\intercal$

to be distributed as $(\mathbf{z}^\mathsf{T}, \mathbf{x}^\mathsf{T})^\mathsf{T}$.

Our goal is to test the null hypothesis

$$H_0 : \quad \mathbf{y}_i \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i \tag{3.1}$$

where $\mathbf{z}_{i\mathcal{S}} = (z_{ij_1}, ..., z_{ij_s})^\mathsf{T}$ is a set of genetic factors to test, identified by the index set $\mathcal{S} = \{j_1, ..., j_s\} \subset \{1, ..., p\}$. Special cases include marginal testing, as in traditional GWAS, where $\mathcal{S} = \{j\}$ for some $j \in \{1, ..., p\}$, or global testing where $\mathcal{S} = \{1, ..., p\}$.

To model the longitudinal trajectory, we assume that y_{ir} is a noisy sample of a smooth underlying function $Y_i(\cdot)$, evaluated at the point t_{ir},

$$y_{ir} = Y_i(t_{ir}) + \epsilon_{ir},$$

where, for each i, $Y_i(\cdot)$ is distributed as $Y(\cdot)$ and $E[Y(t)] = \mu(t)$, $\mathrm{Cov}\{Y(s), Y(t)\} = G(s,t)$, for any $s, t \in \mathcal{T}$, and ϵ_{ir} is a random error independent of $Y_i(\cdot), \mathbf{x}_i$, and $\mathbf{z}_i$, with $E[\epsilon_{ir}] = 0$, and $\mathrm{Var}(\epsilon_{ir}) = \sigma^2$. Furthermore, we assume there is an orthogonal expansion of G,

$$G(s,t) = \sum_{k=1}^{\mathcal{K}} \lambda_k \phi_k(s) \phi_k(t),$$

for $s, t \in \mathcal{T}$, where $\{\phi_k(\cdot)\}_{k=1}^{\mathcal{K}}$ are the eigenfunctions associated with non-negative nonincreasing eigenvalues $\{\lambda_k\}_{k=1}^{\mathcal{K}}$, and $\mathcal{K}$ could be infinity. Then we can express the observed data as a linear combination of the population mean $\mu(\cdot)$ and the eigenfunctions

$$y_{ir} = \mu(t_{ir}) + \sum_{k=1}^{\mathcal{K}} \xi_{ik} \phi_k(t_{ir}) + \epsilon_{ir}. \tag{3.2}$$

where

$$\xi_{ik} = \int_{\mathcal{T}} \{Y_i(t) - \mu(t)\} \phi_k(t) dt$$

are independent random variables with $E[\xi_{ik}] = 0$ and $\mathrm{Var}(\xi_{ik}) = \lambda_k$. Thus, $\mathbf{y}_i$ relates to $\mathbf{x}_i$ and $\mathbf{z}_i$ only through the underlying trajectory $Y_i(\cdot)$ whose randomness is captured by the random coefficients $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$. Thus testing (3.1) is equivalent to testing

$$H_0 : \quad \{\xi_{ik}\}_{k=1}^{\mathcal{K}} \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i$$

However, direct assessment of the association between $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$ and $\mathbf{z}_{i\mathcal{S}}$ is difficult since $\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$ are unobservable and $\mathcal{K}$ could be infinity. To handle the high dimensionality in

$\{\xi_{ik}\}_{k=1}^{\mathcal{K}}$, we note that, despite the fact that the underlying trajectory $Y_i(\cdot)$ is a linear combination of a potentially infinite number of eigenfunctions, we in general need only a finite number to account for nearly all of the variability in $\mathbf{y}$. And because the eigenfunctions are ordered so that the kth eigenfunction explains the kth most variability in $\mathbf{y}$, we can then approximate $Y_i(\cdot)$ using only the first K eigenfunctions

$$Y_i(t) \approx \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t),$$

where $K < \infty$ could be chosen such that the first K directions capture a proportion of the variation at least as large as $\wp \in (0, 1]$. To replace $\{\xi_{ik}\}_{k=1}^{K}$ with observable quantities, we propose to use the so-called BLUP

$$\widetilde{\xi}_{ik} = \lambda_k \boldsymbol{\phi}_{ik}^{\mathsf{T}} \Sigma_{\mathbf{y}_i}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \tag{3.3}$$

where $\boldsymbol{\phi}_{ik} = (\phi_k(t_{i1}), ..., \phi_k(t_{ir_i})^{\mathsf{T}}$, $\boldsymbol{\mu}_i = (\mu(t_{i1}), ..., \mu(t_{ir_i})^{\mathsf{T}}$, and $\Sigma_{\mathbf{y}_i} = \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_i)$ such that $(\Sigma_{\mathbf{y}_i})_{rl} = G(t_{ir}, t_{il}) + \sigma^2\delta_{rl}$ and $\delta_{rl} = I_{\{r=l\}}$. In the PACE method of (Yao et al., 2005), $\widetilde{\xi}_{ik}$ was obtained as $E[\xi_{ik}|\mathbf{y}_i]$ under the assumption that ξ_{ik} and ϵ_{ir} are jointly normal, but we don't require normality here. We simply take $\widetilde{\xi}_{ik}$ as an observable and reasonable approximation to ξ_{ik} even if normality does not hold, as has been argued in (Robinson, 1991; Jiang, 1998).

Thus, we propose to test (3.1) by testing

$$H_0^{\dagger}: \quad \{\widetilde{\xi}_{ik}\}_{k=1}^{K} \perp \mathbf{z}_{i\mathcal{S}} \mid \mathbf{x}_i.$$

Taking note that the association we seek to test is conditional on $\mathbf{x}$, one may construct a test for H_0 by regressing $\widetilde{\boldsymbol{\xi}}_i = (\widetilde{\xi}_1, ..., \widetilde{\xi}_K)^{\mathsf{T}}$ onto $(\mathbf{x}_i, \mathbf{z}_{i\mathcal{S}})$. However, this is only valid if the effect of $\mathbf{x}_i$ on $Y_i(\cdot)$ is captured fully based on the model relating $\mathbf{x}_i$ and $\widetilde{\boldsymbol{\xi}}_i$. To remove the effect of $\mathbf{x}_i$ without imposing a strong assumption on how $\mathbf{x}_i$ affects $Y_i(\cdot)$, we center $\mathbf{z}_{i\mathcal{S}}$ as $\mathbf{z}_{i\mathcal{S}}^* = (z_{ij_1}^*, ..., z_{ij_s}^*)$ where for any j

$$z_{ij}^* = z_{ij} - \mu_{z_j}(\mathbf{x}_i),$$

with $\mu_{z_j}(\mathbf{x}_i) = E[z_j|\mathbf{x}_i]$.

To form the test statistic for H_0, we propose to summarize the overall association between $Y(\cdot)$ and $\mathbf{z}_{\mathcal{S}}$ based on the Frobenius norm of the standardized covariance between

$\widetilde{\boldsymbol{\xi}}_i$ and $\mathbf{z}^*_{i\mathcal{S}}$

$$Q_0 = \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widetilde{\boldsymbol{\xi}}_i \mathbf{z}^{*\mathsf{T}}_{i\mathcal{S}} \right\|_F^2. \tag{3.4}$$

Though Q_0 takes a simple form and can be motivated naturally as an estimated covariance (and can thus be considered model-free), it can also be viewed as a variance component score test statistic similar to those considered previously for other regression models (Commenges and Andersen, 1995; Lin, 1997). Details on the derivation of the variance component score test statistic are given in section 3.2.2.

Both $\widetilde{\boldsymbol{\xi}}_i$ and $\mathbf{z}^*_{i\mathcal{S}}$ involve various nuisance parameters that remain to be estimated. First, under mild regularity conditions which we outline in section 3.6, we can use FPCA to estimate $\mu(\cdot), \lambda_k, \phi_k(\cdot), G(\cdot,\cdot)$, and σ^2 by $\widehat{\mu}(\cdot), \widehat{\lambda}_k, \widehat{\phi}_k(\cdot), \widehat{G}(\cdot,\cdot)$, and $\widehat{\sigma}^2$, respectively, via local linear smoothing as in (Hall et al., 2006; Yao et al., 2005). Subsequently, we can estimate $\widetilde{\xi}_{ik}$ by

$$\widehat{\xi}_{ik} = \widehat{\lambda}_k \widehat{\boldsymbol{\phi}}^{\mathsf{T}}_{ik} \widehat{\Sigma}^{-1}_{\mathbf{y}_i} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i) \tag{3.5}$$

for $\widehat{\boldsymbol{\phi}}_{ik} = (\widehat{\phi}_k(t_{i1}), ..., \widehat{\phi}_k(t_{ir_i})^{\mathsf{T}}, \widehat{\boldsymbol{\mu}}_i = (\widehat{\mu}(t_{i1}), ..., \widehat{\mu}(t_{ir_i})^{\mathsf{T}}$, and $\left(\widehat{\Sigma}_{\mathbf{y}_i}\right)_{rl} = \widehat{G}(t_{ir}, t_{il}) + \widehat{\sigma}^2 \delta_{rl}$. To estimate $\mu_{z_j}(\mathbf{x}_i)$, various approaches can be taken depending on the nature of $\mathbf{x}$. For example, when $\mathbf{x}$ is discrete, $\mu_{z_j}(\mathbf{x}_i)$ can be estimated empirically. With continuous $\mathbf{x}$, we may impose a parametric model with

$$\mu_{z_j}(\mathbf{x}) = g_j(\boldsymbol{\theta}_j, \mathbf{x}) \tag{3.6}$$

and obtain $\bar{z}_j(\mathbf{x})$ as $g_j(\widehat{\boldsymbol{\theta}}_j, \mathbf{x})$, where $\widehat{\boldsymbol{\theta}}_j$ is an estimate of a finite-dimensional parameter $\boldsymbol{\theta}_j$. Finally, based on $\{\widehat{\xi}_{ik}\}_{k=1}^K$ and $\bar{z}_j(\mathbf{x}_i)$, our proposed test statistic is

$$Q = \frac{1}{n} \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \widehat{\xi}_{ik} \widehat{z}^*_{ij} \right)^2 = \left\| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}^{*\mathsf{T}}_{i\mathcal{S}} \right\|_F^2. \tag{3.7}$$

where $\widehat{\mathbf{z}}^*_{i\mathcal{S}} = (\widehat{z}^*_{ij_1}, ..., \widehat{z}^*_{ij_s})^{\mathsf{T}}$ and $\widehat{z}^*_{ij} = z_{ij} - \bar{z}_j(\mathbf{x}_i)$.

3.2.2 Connection to mixed effects models

In this section, we demonstrate that the quantity (3.4) can be arrived at from a more familiar mixed effects model. Consider the model

$$y_{ir} = \mu(t_{ir}) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t_{ir}) + \epsilon_{ir}, \tag{3.8}$$

$$\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{iK})^\intercal \sim N(\mathbf{B}\mathbf{z}^*_{i\mathcal{S}}, \Lambda), \quad \epsilon_{ir} \sim N(0, \sigma^2) \tag{3.9}$$

where $\mathbf{B}$ is a $K \times s$ matrix with (k, j)th entry β_{kj} and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_K)$. We can obtain Q_0 as the variance component score test statistic for $H_0 : \mathbf{B} = 0$. Specifically, let $\beta_{kj} = \eta\nu_{kj}$ and we consider a working model such that $\{\nu_{kj}\}$ are independently distributed with $E(\nu_{kj}) = 0$ and $\text{Var}(\nu_{kj}) = \lambda_k^2$. Under this working model, $H_0 : \mathbf{B} = 0$ is equivalent to

$$H_0 : \eta = 0.$$

To obtain the variance component test statistic, rewrite the model as

$$\mathbf{y}_{\mu i} = \sum_{k=1}^{K} \left(\sum_{j \in \mathcal{S}} \eta\nu_{kj} z^*_{ij} + e_{ik} \right) \boldsymbol{\phi}_{ik} + \boldsymbol{\epsilon}_i$$

for centered outcome $\mathbf{y}_{\mu i} = (y_{i1} - \mu(t_{i1}), ..., y_{ir_i} - \mu(t_{ir_i}))^\intercal$, error vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{ir_i})^\intercal$, and random effects $\mathbf{e}_i = (e_{i1}, ..., e_{iK})^\intercal \sim N(0, \Lambda)$. Then

$$\mathbf{y}_{\mu i} | \, \boldsymbol{\nu}, \{\mathbf{z}^*_{i\mathcal{S}}\}_{i=1}^n \sim N\left(\sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \eta\nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik}, \Sigma_{\mathbf{y}_i} \right)$$

where $\Sigma_{\mathbf{y}_i} = \sum_{k=1}^{K} \lambda_k \boldsymbol{\phi}_{ik} \boldsymbol{\phi}_{ik}^\intercal + \sigma^2 I_{r_i}$ and I_{r_i} is the $r_i \times r_i$ identity matrix.

The likelihood for $\mathbf{y}_{\mu i}$ can then be written as

$$\mathcal{L}(\eta) = \exp\left[\sum_{i=1}^{n} \left\{ -\frac{1}{2} \log |\Sigma_{\mathbf{y}_i}| - \right.\right.$$
$$\left.\left. \frac{1}{2} \left(\mathbf{y}_{\mu i} - \eta \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik} \right)^\intercal \Sigma^{-1}_{\mathbf{y}_i} \left(\mathbf{y}_{\mu i} - \eta \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \nu_{kj} z^*_{ij} \boldsymbol{\phi}_{ik} \right) \right\} \right].$$

Because the target of inference is η, we marginalize over the nuisance parameter $\boldsymbol{\nu}$ conditional on the observed data to obtain $\mathcal{L}^*(\eta) = E[\mathcal{L}(\eta)|\mathbb{V}]$ where the expectation is

taken over the distribution of $\boldsymbol{\nu}$. We follow the argument in Commenges and Andersen (1995) and note that the score at the null value is 0: $\lim_{\eta\to 0}\partial \log \mathcal{L}^*(\eta)/\partial\eta = E\left[\sum_{i=1}^n \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^K \nu_{kj} z_{ij}^* \boldsymbol{\phi}_{ik} \mid \mathbb{V}\right] = 0$. So we instead consider the score with respect to $\eta^2, \lim_{\eta\to 0}\partial \log \mathcal{L}^*(\eta)/\partial(\eta^2)$, and we show in section 3.6.3 that it can be written as

$$
\begin{aligned}
&E\left[\left.\frac{\partial \log \mathcal{L}(\eta)}{\partial \eta}\right|_{\eta=0} \mid \mathbb{V}\right]^2 + E\left[\left.\frac{\partial^2 \log \mathcal{L}(\eta)}{\partial \eta^2}\right|_{\eta=0} \mid \mathbb{V}\right] \\
&= E\left[\sum_{i=1}^n \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^K \nu_{kj} z_{ij}^* \boldsymbol{\phi}_{ik} \mid \mathbb{V}\right]^2 - \\
&E\left[\sum_{i=1}^n \sum_{j,j'\in\mathcal{S}} \sum_{k,k'=1}^K \nu_{kj} z_{ij}^* \boldsymbol{\phi}_{ik}^\top \Sigma_{\mathbf{y}_i}^{-1} \nu_{k'j'} z_{ij'}^* \boldsymbol{\phi}_{ik'} \mid \mathbb{V}\right] \\
&= E\left[\sum_{j\in\mathcal{S}} \sum_{k=1}^K \nu_{kj} \left(\sum_{i=1}^n \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} z_{ij}^* \boldsymbol{\phi}_{ik}\right) \mid \mathbb{V}\right]^2 - \\
&\quad E\left[\sum_{j\in\mathcal{S}} \sum_{k=1}^K \nu_{kj}^2 \left\{\sum_{i=1}^n (z_{ij}^*)^2 \boldsymbol{\phi}_{ik}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik}\right\} \mid \mathbb{V}\right] \\
&= \sum_{j\in\mathcal{S}} \sum_{k=1}^K \left(\sum_{i=1}^n \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k z_{ij}^*\right)^2 - \\
&\quad \sum_{j\in\mathcal{S}} \sum_{k=1}^K \left\{\sum_{i=1}^n (\lambda_k z_{ij}^*)^2 \boldsymbol{\phi}_{ik}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik}\right\}
\end{aligned}
$$

up to a scaling constant.

To finally obtain Q_0, we standardize by n^{-1} and drop the second term because it converges to a constant, yielding the score statistic

$$
\begin{aligned}
&n^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^K \left(\sum_{i=1}^n \mathbf{y}_{\mu i}^\top \Sigma_{\mathbf{y}_i}^{-1} \boldsymbol{\phi}_{ik} \lambda_k z_{ij}^*\right)^2 \\
&= n^{-1} \sum_{j\in\mathcal{S}} \sum_{k=1}^K \left(\sum_{i=1}^n \widetilde{\xi}_{ik} z_{ij}^*\right)^2 = Q_0.
\end{aligned}
$$

taking note of the form of $\widetilde{\xi}_{ik}$ from (3.3). Thus, our proposed test statistic can be obtained as a variance component test under a normal mixed model framework. On the other hand, we can also view Q_0 as a simple summary of the overall covariance between the FPCA scores

and the genetic markers. We next derive the null distribution of the FPVC test statistic without requiring the normal mixed model to hold.

3.2.3 Estimating the null distribution of the test statistic

In order to obtain p-values for FPVC testing, we must identify the null distribution of Q. To this end, we show in section 3.6.2 that the key quantity in Q

$$q_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\xi}_{ik} \widehat{z}^*_{ij}$$

is asymptotically equivalent to

$$\widetilde{q}_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widetilde{\xi}_{ik} \widehat{z}^*_{ij}$$

under H_0, i.e. $q_{kj} - \widetilde{q}_{kj} = o_p(1)$ for each j and k. The key idea for deriving the null distribution of q_{kj} is that, since $\widehat{z}^*_{ij}$ is approximately mean 0 conditional on $\mathbf{x}_i$, the variability due to approximating $\widetilde{\xi}_{ik}$ by $\widehat{\xi}_{ik}$ does not contribute any additional noise to q_{kj} (compared to $\widetilde{q}_{kj}$) at the first order under H_0. Thus, we can obtain the limiting distribution of Q by analyzing the quantity $\widetilde{Q} = \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \widetilde{q}^2_{kj}$.

To characterize the null distribution of $\widetilde{Q}$, we need to account for the variability in the estimated model parameters for $\mu_{z_j}(\mathbf{x}_i) = g_j(\boldsymbol{\theta}_j, \mathbf{x}_i)$ in $\widehat{z}^*_{ij}$. Without loss of generality, we assume that for each j

$$n^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{U}_j(\mathbf{x}_i) z^*_{ij} + o_p(1), \tag{3.10}$$

where $\mathcal{U}(\cdot)$ is some $(q+1)$-dimensional function of $\mathbf{x}_i$ with $E|\mathcal{U}(\mathbf{x}_i)| < \infty$. It follows that

$$q_{kj} = \widetilde{q}_{kj} + o_p(1) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{Q}_{ikj} + o_p(1) \tag{3.11}$$

where $\mathcal{Q}_{ikj} = \left\{ \widetilde{\xi}_{ik} - \mathbb{A}_{kj} \mathcal{U}(\mathbf{x}_i) \right\} z^*_{ij}$, $\mathbb{A}_{kj} = E[\widetilde{\xi}_{ik} \dot{\boldsymbol{g}}_j(\boldsymbol{\theta}_j, \mathbf{x}_i)^\intercal]$, and $\dot{\boldsymbol{g}}_j(\boldsymbol{\theta}_j, \mathbf{x}_i) = \partial g_j(\boldsymbol{\theta}_j, \mathbf{x}_i)/\partial \boldsymbol{\theta}_j$. We show in the appendix that the limiting null distribution of Q is a mixture of χ^2_1 random variables, $Q \sim \sum_{l=1}^{sK} a_l \chi^2_1$, with mixing coefficients determined by the eigenvalues of the variance covariance matrix of $\{\mathcal{Q}_{ikj}\}_{j \in \mathcal{S}, 1 \le k \le K}$. So finally we obtain a p-value for the association between the set $\mathbf{z}_\mathcal{S}$ and $Y(\cdot)$ as $P(\sum_{l=1}^{sK} \widehat{a}_l \chi^2_1 > Q \mid \mathbb{V})$, where $\widehat{a}_l$ is an empirical estimate of a_l.

By a similar argument, one could construct an asymptotically equivalent test statistic by estimating $\widetilde{\boldsymbol{\xi}}_i$ in two stages. Instead of obtaining an estimator directly from FPCA via equation (3.5), FPCA can be used to estimate only $\mu(\cdot)$ and $\{\phi_k(\cdot)\}_{k=1}^K$. By plugging the estimated $\widehat{\mu}(\cdot)$ and $\{\widehat{\phi}_k(\cdot)\}_{k=1}^K$ into the mixed model (3.8), one can obtain what we will call the *re-fitted* test statistic

$$\bar{Q} = n^{-1} \sum_{j \in \mathcal{S}} \sum_{k=1}^{K} \left[\sum_{i=1}^{n} \bar{\xi}_{ik} \widehat{z}_{ij}^{*} \right]^2 \tag{3.12}$$

where $\bar{\boldsymbol{\xi}}_i = (\bar{\xi}_{i1}, ..., \bar{\xi}_{iK})^\intercal$ is the BLUP from the model $y_{ir} - \widehat{\mu}(t_{ir}) = \sum_{k=1}^{K} \xi_{ik} \widehat{\phi}_k(t_{ir}) + \epsilon_{ir}$ with $\text{Cov}(\boldsymbol{\xi}_i) = D$, for some unspecified positive definite matrix D. By the same argument above, estimation of $\widetilde{\xi}_{ik}$ by $\bar{\xi}_{ik}$ contributes no additional variability to the test statistic at the first order. It follows that

$$q_{kj}^{\dagger} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \bar{\xi}_{ik} \widehat{z}_{ij}^{*} = \widetilde{q}_{kj} + o_p(1).$$

and hence $\bar{Q}$ has the same limiting null distribution as Q. Not surprisingly, simulation results suggest that the performance of $\bar{Q}$ is quite similar to the performance of Q. This equivalence indicates that effectively our proposed testing procedure uses FPCA to estimate potentially nonlinear bases and assesses the effect of genetic markers by fitting a mixed model with these basis functions. On the other hand, the test statistics can also be viewed as a simple summary of covariances, and – since we estimate the null distribution without relying on the normality assumption required by the mixed models – our testing procedure remains valid regardless of the adequacy of the mixed model.

3.2.4 Combining multiple sources of outcome information

For settings where disease progression can be better characterized by trajectories of multiple markers, it would be desirable to test the overall association between the genetic factors and all available markers. For example HIV progression is often characterized by both $CD4^+$ cell count and viral load (among other measures) over time. FPVC testing can be easily adapted to perform a test for the overall association between $\mathbf{z}_{\mathcal{S}}$ and all outcomes of interest. To use information in multiple outcomes, $\{\mathbf{y}^{(m)}\}_{m=1}^{M}$, we simply perform FPCA separately on each $\mathbf{y}^{(m)}$ and obtain FPCA scores for each person and each outcome. Subject i's scores for $\mathbf{y}_i^{(m)}$

would be $\widehat{\boldsymbol{\xi}}_i^{(m)} = (\widehat{\xi}_{i1}^{(m)}, ..., \widehat{\xi}_{iK_m}^{(m)})^\intercal$, as in (3.5), and the full set of scores for person i would be $\widehat{\boldsymbol{\xi}}_i = (\widehat{\boldsymbol{\xi}}_i^{(1)\intercal}, ..., \widehat{\boldsymbol{\xi}}_i^{(m)\intercal})^\intercal$. Then we simply proceed by testing

$$H_0 : \{\mathbf{y}^{(m)}\}_{m=1}^M \perp \mathbf{z}_\mathcal{S} \mid \mathbf{x}$$

as before based on

$$Q = \|n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}_{i\mathcal{S}}^{*\intercal}\|_F^2 = \sum_{m=1}^M \|n^{-\frac{1}{2}} \sum_{i=1}^n \widehat{\boldsymbol{\xi}}_i^{(m)} \widehat{\mathbf{z}}_{i\mathcal{S}}^{*\intercal}\|_F^2. \tag{3.13}$$

Since each outcome may be measured on a different scale, one may use scaling or weighting to allow scores from each outcome to contribute similarly to the test statistic. See section 3.5 for further discussion of scaling/weighting.

3.3 Simulation results

We have performed simulation studies to assess the finite sample performance of our proposed testing procedure and compare its power to the standard linear-mixed-model-based procedures. For simplicity, we focused on a single marker z in the absence of covariates and two potential functional outcomes generated from

$$\begin{aligned} y_{ir}^{(m)} &= Y_i^{(m)}(t_{ir}) + \epsilon_{ir}^{(m)} \\ &= \sin(t_{ir}) + (-1)^{m-1}\gamma \left\{\sin(t_{ir}/3) + \cos(t_{ir})\right\} + \\ &(1-\gamma)\left\{b_{i0} + 0.5 b_{i1}^{(m)} \cos(t_{ir}/4)\right\} + \\ &\qquad \beta z_i \left\{\alpha \left(\cos(t_{ir}) + \cos(t_{ir}/10) - \sin(3t_{ir})\right) + \right. \\ &\left. (1-\alpha) t_{ir}/7\right\} + \epsilon_{ir}^{(m)}, \qquad m = 1, 2, \end{aligned}$$

where $b_{ij}^{(m)} \sim N(0, 0.25), j = 0, 1$ are independent and identically distributed (iid) random effects and $\epsilon_{ir}^{(m)} \sim N(0, 0.25)$ are iid errors, for $m = 1, 2$. For each subject i, we generate the number of observations from a Poisson distribution $r_i \sim \text{Poisson}(6) + 2$, and we generate t_{ir} uniformly over the time interval $(0, 2\pi)$. The parameter β controls the magnitude of the genetic effect. The parameter α controls how linear the genetic affect is – when $\alpha = 0$ the genetic effect is entirely linear, and when $\alpha = 1$ the effect is entirely nonlinear. On the

other hand, γ controls the complexity of the mean process and the amount of inter-subject variability – when $\gamma = 0$, the mean process is relatively simple but the inter-subject variability is high, and when $\gamma = 1$ the mean process is complex and the inter-subject variability is low. The genetic factor z_i is generated according to a binomial(2, MAF), with MAF the minor allele frequency.

We examined the performance of the FPVC test statistic Q (defined in (3.13), here denoted by "FPCA") and its asymptotically equivalent counterpart $\bar{Q}$ (defined in the context of a single outcome in (3.12), here denoted "Re-fitted"). For the purposes of comparison, we also examined the performance of a similar test statistic that does not use FPVC but instead employs a pre-specified basis. Consider the test statistic $Q_{\text{lin}} = \frac{1}{n}\sum_{m=1}^{2}\sum_{k=1}^{2}\left[\sum_{i=1}^{n}\xi_{ik}^{(m)\dagger}\widehat{z}_i^*\right]^2$, where $\xi_{ik}^{(m)\dagger}$ is the BLUP from the linear mixed model $y_{ir} = \beta_0 + \beta_1 t_{ir} + \xi_{i1} + \xi_{i2} t_{ir} + \epsilon_{ir}$. In the following, we denote results for Q_{lin} by "Linear".

The number of FPCA scores for the mth outcome, K_m, was selected as the smallest K such that the fraction of variation explained (FVE), $\sum_{k=1}^{K}\widehat{\lambda}_k/(\sum_k \widehat{\lambda}_k)$, was at least $\wp = 0.99$. In the following we report power as the proportion of 1000 simulations for which the testing procedure produced a p-value below 0.05. To ensure that the scores for each outcome contributed comparably to the test statistics, we centered and scaled each outcome as $y_{ir}^{*(m)} = (y_{ir}^{(m)} - \bar{y}^{(m)})/\widehat{\sigma}_y^{(m)}$, prior to obtaining $\widehat{\xi}_{ik}^{(m)}$ and $\xi_{ik}^{(m)\dagger}$, where $\widehat{\sigma}_y^{(m)} = \sqrt{(n-1)^{-1}\sum_{i,r}(y_{ir}^{(m)} - \bar{y}^{(m)})^2}$ and $\bar{y}^{(m)} = n^{-1}\sum_{i,r} y_{ir}^{(m)}$.

To demonstrate the role of sample size and MAF, we simulated data at sample sizes $n = 200, 400$, and 600, and at MAFs 0.1, 0.2, and 0.3. Figure 3.1 displays results for $\gamma = 0$ and $\alpha = 1$. The figure indicates that the performances of Q and $\bar{Q}$ was similar and uniformly dominated the performance of Q_{lin} at any sample size and MAF. Each test maintained its nominal size, with empirical type I error rates for Q and $\bar{Q}$ ranging between 0.034 (MAF = 0.2, $n = 400$) and 0.057 (MAF = 0.2 and $n = 600$ for Q and MAF = 0.1 and $n = 400$ for $\bar{Q}$) and for Q_{lin} between 0.036 (MAF = 0.2 and $n = 400$) and 0.055 (MAF = 0.1 and $n = 600$).

By varying α, we investigated the performance of each method for various levels of linearity in the genetic effect over time. Figure 3.2 displays results for $n = 200$, MAF = 0.1, $\gamma = 1$, and varying α. The figure demonstrates that, despite the fact that the true effect

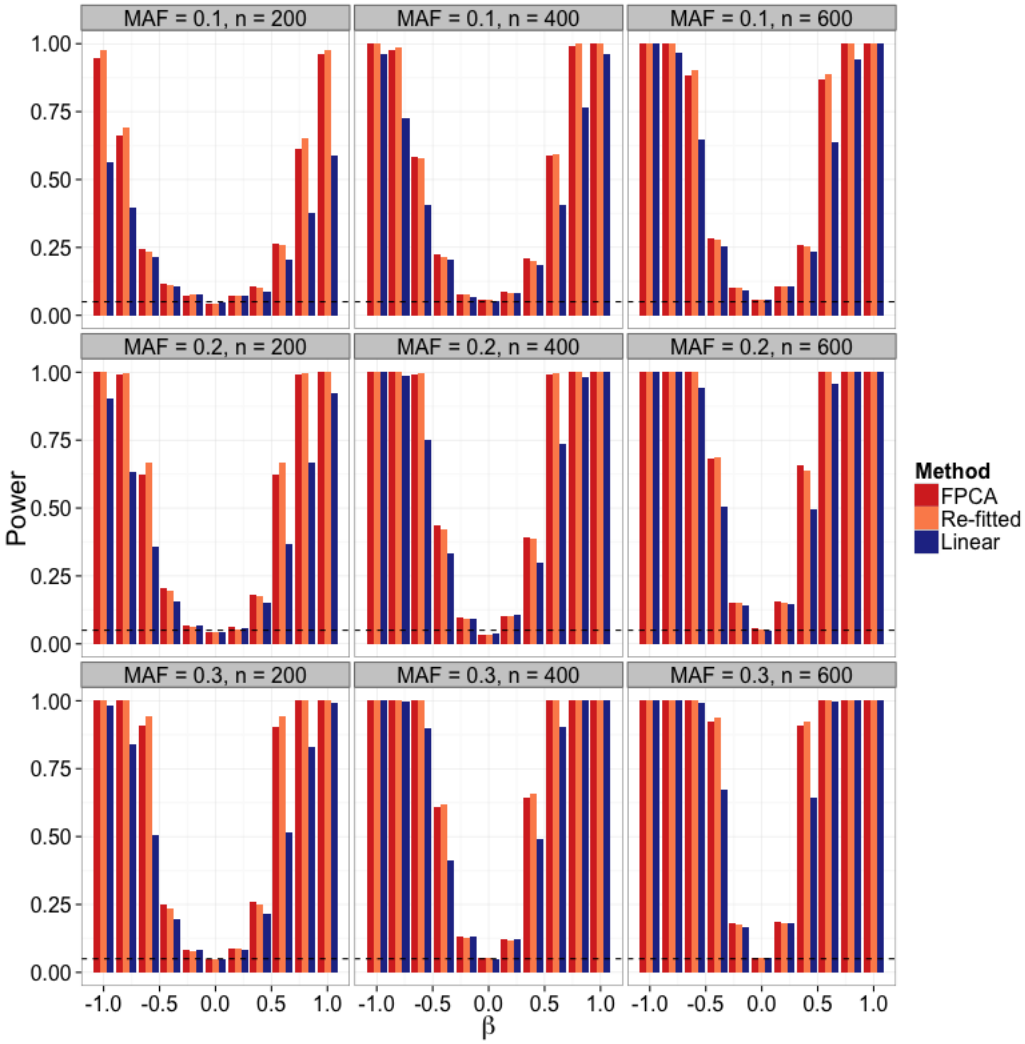


Figure 3.1: Power to detect β using Q (FPCA), $\bar{Q}$ (Re-fitted), and Q_{lin} (Linear), with simple mean process and high inter-subject variability ($\gamma = 0$) and linear genetic effect ($\alpha = 1$).

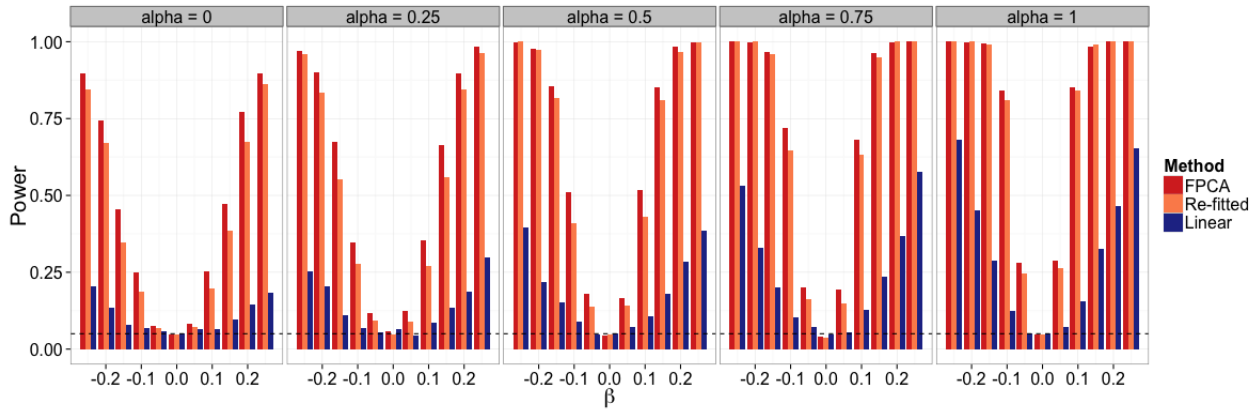


Figure 3.2: Power to detect β using Q (FPCA), $\bar{Q}$ (Re-fitted), and Q_{lin} (Linear), with complex mean process and low inter-subject variability ($\gamma = 1$), sample size $n = 200$, and MAF $= 0.1$. Panels correspond to varying linearity of genetic effect (varying levels of α).

was linear when $\alpha = 0$, the advantage in terms of power of using FPVC did not diminish as α decreased from 1 to 0. Again, the empirical type I error rates hovered around the desired value of 0.05: ranging from 0.041 ($\alpha = 0.75$) to 0.057 ($\alpha = 0.25$) for Q; from 0.037 ($\alpha = 0.75$) to 0.048 ($\alpha = 0$) for $\bar{Q}$; and from 0.046 ($\alpha = 0.75$) to 0.065 ($\alpha = 0.25$) for Q_{lin}.

Similarly, as we varied γ, we saw power gains by using the FPVC-based Q and $\bar{Q}$, with the gains increasing as γ approached 1, the functional form of $Y_i^{(m)}(\cdot)$ became more complex, and the need to flexibly model it increased. Figure 3.3 demonstrates this for $n = 200$, MAF $= 0.1$, $\alpha = 1$, and varying γ. The empirical type I error rates ranged from 0.040 ($\gamma = 0.25$) to 0.048 ($\gamma = 0$) for Q; from 0.036 ($\gamma = 1$) to 0.047 ($\gamma = 0$) for $\bar{Q}$; and from 0.040 ($\gamma = 0.75$) to 0.059 ($\gamma = 1$) for Q_{lin}.

In all of our simulations, the FPVC methods dominated the linear method in terms of power while maintaining desirable type I error rates. We wanted to ensure that the improvement we were seeing was not simply due to the fact that the linear model used only two scores, a random intercept $\xi_{i1}^{(m)\dagger}$ and a random slope $\xi_{i2}^{(m)\dagger}$, for each outcome whereas the FPVC-based methods used K_m scores, where K_m was often selected larger than 2. Thus, we also considered the performance of the FPVC tests when the number of scores was fixed at 2, 3, and 4, and we compared to similar tests based on pre-specified bases with the number of scores fixed at the same respective number.

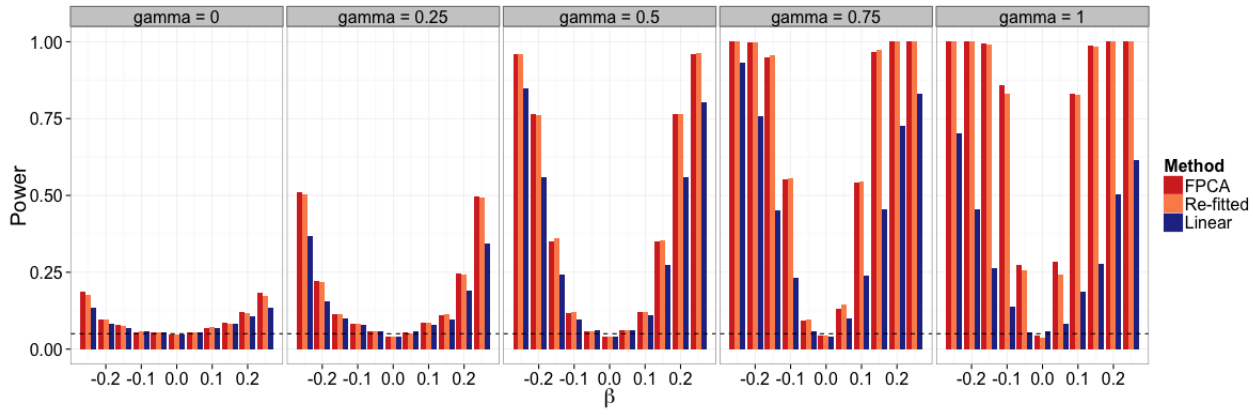


Figure 3.3: Power to detect β using Q (FPCA), $\bar{Q}$ (Re-fitted), and Q_{lin} (Linear), with linear genetic effect $\alpha = 1$, $n = 200$, and MAF $= 0.1$. Panels correspond to varying levels of complexity in the mean structure and varying levels of inter-subject variability (varying levels of γ).

Specifically, we compared the FPVC test statistics Q and $\bar{Q}$ with $K_m = 2$ to Q_{lin}; we compared the FPVC test statistics with $K_m = 3$ to $Q_{\text{quad}} = \frac{1}{n}\sum_{m=1}^{2}\sum_{k=1}^{3}\left[\sum_{i=1}^{n}\xi_{ik}^{(m)\dagger}\widehat{z}_i^*\right]^2$, where $\xi_{ik}^{(m)\dagger}$ is the BLUP from the mixed model $y_{ir} = \beta_0 + \beta_1 t_{ir} + \beta_2 t_{ir}^2 + \xi_{i1} + \xi_{i2} t_{ir} + \xi_{i3} t_{ir}^2 + \epsilon_{ir}$; and we compared the FPVC test statistics with $K_m = 4$ to $Q_{\text{cube}} = \frac{1}{n}\sum_{m=1}^{2}\sum_{k=1}^{4}\left[\sum_{i=1}^{n}\xi_{ik}^{(m)\dagger}\widehat{z}_i^*\right]^2$, where $\xi_{ik}^{(m)\dagger}$ is the BLUP from the mixed model $y_{ir} = \beta_0 + \beta_1 t_{ir} + \beta_2 t_{ir}^2 + \beta_3 t_{ir}^3 + \xi_{i1} + \xi_{i2} t_{ir} + \xi_{i3} t_{ir}^2 + \xi_{i4} t_{ir}^3 + \epsilon_{ir}$.

We found that there were some situations when using the pre-specified polynomial basis could outperform the FPVC tests, particularly when γ was near 0 and α was near 1. Even in the cases when it was possible to outperform the FPVC-based methods with a pre-specified basis, though, one would have to know the correct number of scores to use *a priori*. For example, as in figure 3.4, when all methods used 2 scores, the FPVC-based methods saw large increases in power over the pre-specified model (empirical type I error rates: 0.049 for Q and $\bar{Q}$, 0.052 for Q_{lin}), and when all methods used 4 scores the FPVC-based methods performed slightly better (empirical type I error rates: 0.050 for Q, 0.047 for $\bar{Q}$, 0.049 for Q_{cube}). However, when all methods used 3 scores, the pre-specified model saw a large increase in power over the FPVC-based methods (empirical type I error rates: 0.050 for Q and $\bar{Q}$, 0.058 for Q_{quad}). This suggests that, in practice, if the correct basis were known or known approximately, then a pre-specified basis may be able to obtain higher power. However,

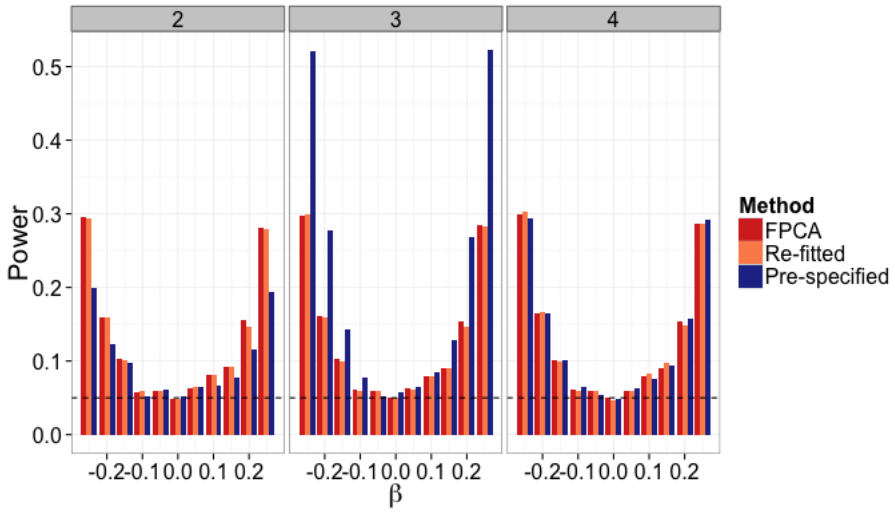


Figure 3.4: Power to detect β using Q (FPCA), $\bar{Q}$ (Re-fitted), and pre-specified basis test statistics $Q_{\text{lin}}, Q_{\text{quad}}, Q_{\text{cube}}$ (Pre-specified), with mostly nonlinear genetic effect ($\alpha = 0.75$), and low complexity mean process and high inter-subject variability ($\gamma = 0.25$), sample size $n = 200$, and MAF $= 0.1$. Panels indicate how many scores were used in testing.

over- or under-specification of the model may result in decreased performance. With FPVC testing, on the other hand, the number of scores chosen does not have a dramatic effect on performance. Thus, using FPVC alleviates the need to know precisely how many basis functions to use.

Furthermore, there were many simulation settings where the pre-specified model never approached the FPVC tests. Consider when $\alpha = 0$ and $\gamma = 0.75$, as in figure 3.5. There, even if you increased the number of scores used to 4, the FPVC-based methods were much more powerful than those using a pre-specified basis.

3.4 Association between genetics and trajectory of LDL

We applied our method to a study of the association between longitudinal LDL cholesterol and a set of candidate single-nucleotide polymporphisms (SNPs). The study cohort consisted of 2840 RA cases and controls identified via electronic health records (EHR) (Liao et al., 2010). LDL measurements were also obtained via EHR. A total of 26 candidate SNPs were identified as potentially associated with LDL. Here we are interested in assessing the

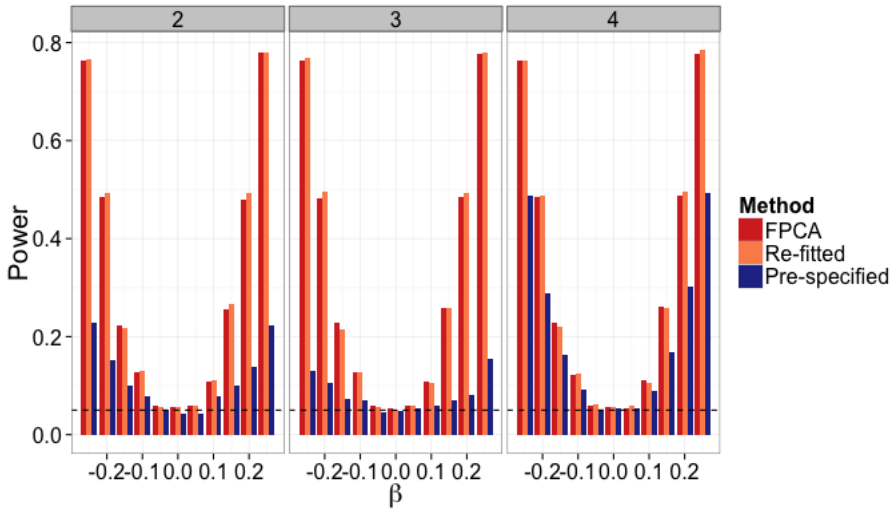
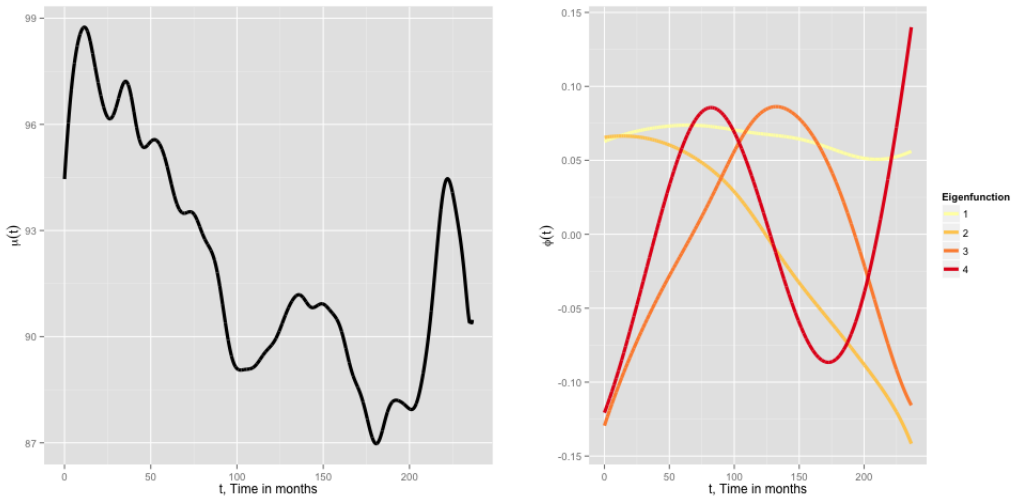


Figure 3.5: Power to detect β using Q (FPCA), $\bar{Q}$ (Re-fitted), and pre-specified basis test statistics $Q_{\text{lin}}, Q_{\text{quad}}, Q_{\text{cube}}$ (Pre-specified), with linear genetic effect ($\alpha = 0$), high complexity mean process and low inter-subject variability ($\gamma = 0.75$), sample size $n = 200$, and MAF $= 0.1$. Panels indicate how many scores were used in testing.

association between these risk alleles and the trajectory of LDL over time.

FPCA was performed on all 2840 individuals with any LDL observations, but testing was performed on only those 1901 with genetic information. Patients contributed between 1 and 93 LDL measurements, with the median being 19, and approximately 90% of patients contributed between 4 and 50 LDL observations over time. Patients with only 1 measurement contributed to the estimation of the mean function $\mu(\cdot)$ but not to the estimation of the eigenfunctions. The estimated mean function and eigenfunctions for LDL are depicted in figure 3.6. The eigenfunctions roughly approximate a polynomial basis, with the first eigenfunction approximately constant, the second approximately linear, the third approximately quadratic, and the fourth approximately cubic.

Because the sampling mechanism of the study cohort depended on RA status, we analyzed the association between LDL and the SNPs separately in RA cases and controls. We combined the resulting p-values for cases and controls using Fisher's method. We first performed a global test of any association between the entire SNP set $(z_1, ..., z_{26})^\intercal$ and the trajectory of LDL, yielding a p-value of 0.21 in cases and 0.17 in controls, for an overall p-value of 0.15. Thus, there was not significant evidence of any association. We next proceeded



(a) Mean (b) Eigenfunction

Figure 3.6: FPCA results. Estimated (a) population mean and (b) eigenfunctions. In plot (b), color denotes eigenfunction.

with exploratory analysis of the individual SNPs.

The results for marginal testing of all SNPs are depicted in figure 3.7. Five SNPs obtained p-values below 0.10, and two of those were below 0.05. The lowest p-value ($p = 0.032$) belonged to rs2081687 near the CYP7A1 gene, which had previously been associated with non-longitudinal LDL in a population of European ancestry and replicated in a population of African Americans (Adeyemo et al., 2012). The other SNP with a p-value below 0.05 was rs1564348 ($p = 0.046$) which is near the gene LPA and had previously exhibited an association with decreased levels of non-longitudinal apolipoprotein(a) and LDL (Holmes et al., 2011).

In order to better visualize the differences FPVC testing was picking up on, we plot the mean estimated trajectories of LDL among those with 0, 1, and 2 minor alleles, respectively, in figure 3.8 for the SNP rs2081687. We obtained the mean estimated trajectories by averaging the scores $\widehat{\xi}_{ik}$ across all individuals with the same number of minor alleles. In both cases and controls, those with 0 minor alleles have the uniformly lowest estimated trajectory, and those with 2 minor alleles have the uniformly highest estimated trajectory. In RA cases, those with 1 and 2 minor alleles have relatively similar trajectories, while in controls all three curves show quite a bit of separation. In both cases and controls, separation between the

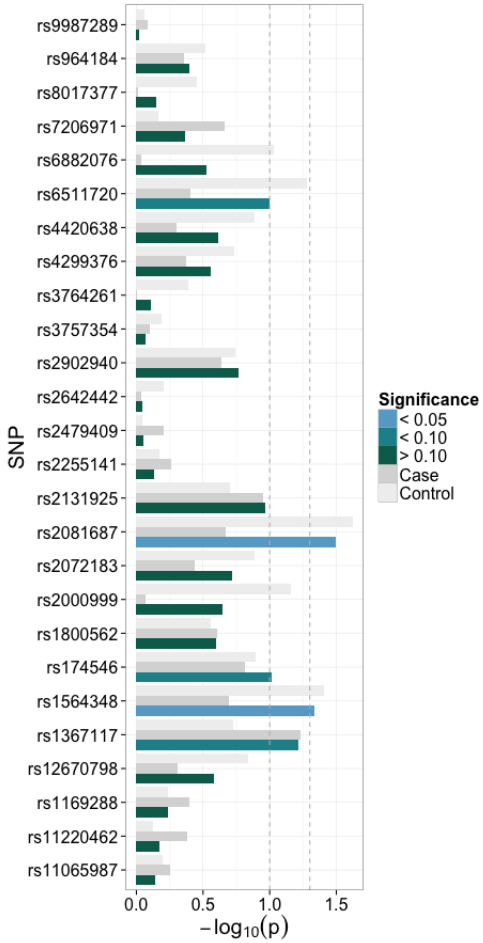


Figure 3.7: P-values of association between longitudinal LDL and 26 candidate SNPs. Colored bars represent p-values of overall association, combining information from cases and controls. Gray bars indicate component p-values computed either only in cases or only in controls. P-values are given on the -$\log_{10}$ scale. For overall p-values, color denotes $p < 0.05$ (blue), $p < 0.10$ (teal), $p > 0.10$ (dark green). For component p-values, color denotes whether it was obtained from cases (dark gray) or controls (light gray).

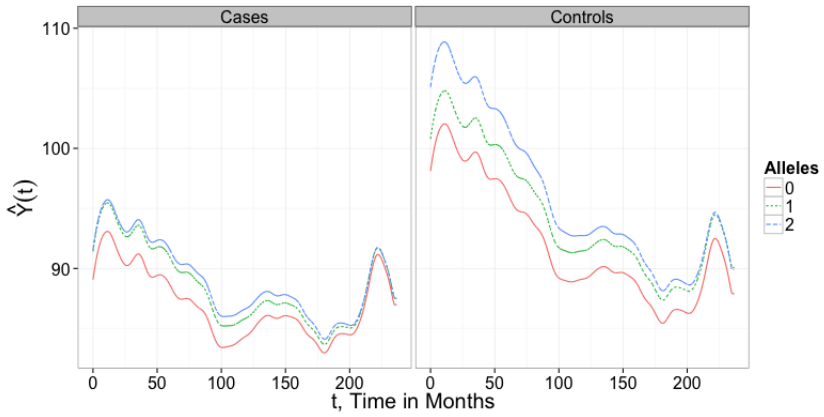


Figure 3.8: Estimated mean trajectories among patients with a given number of minor alleles at rs2081687. Color and line type denote number of minor alleles. Cases are depicted in the left panel, and controls are depicted in the right panel.

curves seems to be largest at the earliest time points.

3.5 Discussion

We have proposed functional principal variance component testing, a FPCA-based testing procedure for assessing the association between a set of genetic variants $\mathbf{z}_{\mathcal{S}}$ and a complexly varying longitudinal outcome $\mathbf{y}$ that is feasible on the genome-wide scale, allowing adjustment for other covariates. Unlike the standard mixed-model-based approaches, we do not model the trajectories $\{Y_i(\cdot)\}_{i=1}^n$ parametrically but use the data to identify the most parsimonious summaries of the trajectory patterns via FPCA. We subsequently test the association between the random coefficients $\boldsymbol{\xi}_i$ and the markers of interest using a test statistic motivated by variance component testing. Our procedure could potentially be much more powerful than procedures based on pre-specified bases, which might suffer power loss due to either high degrees of freedom or inability to capture the complexity in the trajectories. Furthermore, our FPVC testing is computationally efficient as we are able to perform thousands or even millions of tests quickly by separating the time-intensive FPCA from the testing. This makes our method feasible on the genome-wide scale where millions of marginal tests may be necessary.

It is important to note that while we make mild assumptions on the longitudinal outcome

$\mathbf{y}$ to obtain the form of our proposed test statistic, the validity of FPVC testing requires no assumption about the relationship between $\mathbf{y}$ and $\mathbf{z}_{\mathcal{S}}$. FPVC testing remains valid even if the working mixed model (3.8) fails to hold. Additionally, while one can motivate the quantity $\widetilde{\xi}_{ik}$ as the conditional expectation of ξ_{ik} under a normality assumption on ξ_{ik} and ϵ_{ir}, even when this normality fails to hold, testing based on Q remains valid since the estimated eigenvalues and eigenfunctions from functional PCA are uniformly convergent to their limits (Hall et al., 2006). In fact, one can consider FPVC model-free in that the test statistic Q could be motivated simply as an estimated covariance. Furthermore, we assume that the errors ϵ_{ir} are iid with mean 0 and variance σ^2, but some relaxation of this assumption is possible for some "degree of weak dependence and in cases of non-identical distribution" (Hall et al., 2006), while still maintaining the validity of our procedure.

FPVC testing can also simultaneously consider multiple sources of outcome information to better characterize complex phenotypes. With multiple longitudinal outcomes, one might wish to ensure that scores for all outcomes are roughly on the same scale, so that each outcome contributes comparably to the test statistic. To this end, one may consider a weighted version of (3.13) as

$$Q = \sum_{m=1}^{M} \omega_m \| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i \widehat{\mathbf{z}}_{i\mathcal{S}}^{*} \|_F^2,$$

where ω_m are nonnegative outcome-specific weights that can be pre-specified or data adaptive. Alternatively, in the absence of relevant weights, one can simply scale each $\mathbf{y}^{(m)}$ so that the magnitude of $\widehat{\boldsymbol{\xi}}_i^{(m)}$ is comparable across different values of m. Let $y_{ir}^{*(m)} = y_{ir}^{(m)} / \widehat{\sigma}_y^{(m)}$ where $\widehat{\sigma}_y^{(m)} = \sqrt{(n-1)^{-1} \sum_{i,r} (y_{ir}^{(m)} - \bar{y}^{(m)})^2}$ and $\bar{y}^{(m)} = n^{-1} \sum_{i,r} y_{ir}^{(m)}$. Then obtain $\widehat{\boldsymbol{\xi}}_i^{*(m)}$ via FPCA on $\{\mathbf{y}_i^{*(m)}\}_{i=1}^n$ and construct the test statistic $\sum_{m=1}^{M} \| n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\boldsymbol{\xi}}_i^{*(m)} \widehat{\mathbf{z}}_{i\mathcal{S}}^{*} \|_F^2$. Such a strategy appears to work well in simulation studies.

3.6 Appendix

3.6.1 FPCA Assumptions

The primary result regarding FPCA we require is that the estimated scores $\widehat{\xi}_{ik}$ converge in probability to the BLUP, $\widehat{\xi}_{ik} \longrightarrow_p \widetilde{\xi}_{ik}$.

We reproduce here the assumptions from Hall et al. (2006) for this result, with minor modification.

A.1 There is an integer $i_0 > 1$ for which there are no ties among the i_0+1 largest eigenvalues of $G(\cdot,\cdot)$.

A.2 For the data pairs $(t_{ir}, y_{ir})_{1\le r\le r_i; 1\le i\le n}$, $r_i \ge 2$ and $R_n = \max_{i\le n} r_i$ is bounded by $R < \infty$ as $n \to \infty$.

A.3 The time points t_{ir} have common density f_t, which is bounded away from 0 on $\mathcal{T}$, and are independent of $\mathbf{z}_i$ and $\mathbf{x}_i$.

There are further requirements in Hall et al. (2006) that pertain specifically to smoothing. For example, in the local linear smoother used there, they require assumptions about the second derivatives of the smooth functions $Y_i(\cdot)$ in addition to assumptions about the kernel and bandwidth used for smoothing. Since the mechanics of FPCA are not the focus of this paper, we do not reproduce these conditions here.

To facilitate subsequent proofs, we construct the R-dimensional vectors $\mathbf{t}_i^* = (t_{i1}^*, ..., t_{iR}^*)^\intercal$ and $\mathbf{y}_i^* = (y_{i1}^*, ..., y_{iR}^*)^\intercal$. Define $\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{iR})^\intercal$ to be a *non-missingness indicator* so that $\{t_{ir}, 1 \le r \le r_i\} = \{t_{ir}^* : \delta_r = 1, 1 \le r \le R\}$, $\{y_{ir}, 1 \le r \le r_i\} = \{y_{ir}^* : \delta_r = 1, 1 \le r \le R\}$, and $\delta_{ir} = 0$ for all r such that (t_{ir}^*, y_{ir}^*) is not observed. We finally require that

A.4 For each $1 \le r \le R$, $D_{rn} = \sum_{i=1}^n \delta_{ir} \to \infty$ as $n \to \infty$.

3.6.2 Justification for the asymptotic null distribution

In this section, we establish the limiting null distribution of Q. First, note the form of Q and $\widetilde{Q}$:

$$Q = \mathbf{q}^\intercal \mathbf{q}, \widetilde{Q} = \widetilde{\mathbf{q}}^\intercal \widetilde{\mathbf{q}} \qquad \mathbf{q} = (q_{kj})_{j\in\mathcal{S}, 1\le k\le K}, \widetilde{\mathbf{q}} = (\widetilde{q}_{kj})_{j\in\mathcal{S}, 1\le k\le K}.$$

We show that $q_{kj} - \widetilde{q}_{kj} = o_p(1)$, which ensures that Q and $\widetilde{Q}$ are asymptotically equivalent. Define $\Delta_{kj} = q_{kj} - \widetilde{q}_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} (\widehat{\xi}_{ik} - \widetilde{\xi}_{ik}) \widehat{z}^*_{ij}$ and furthermore that

$$\begin{aligned} \widetilde{\xi}_{ik} &= \lambda_k \phi_k(\mathbf{t}_i)^\intercal \Sigma(\mathbf{t}_i)^{-1} \left\{ \mathbf{y}_i - \mu(\mathbf{t}_i) \right\} \qquad \text{and} \\ \widehat{\xi}_{ik} &= \widehat{\lambda}_k \widehat{\phi}_k(\mathbf{t}_i)^\intercal \widehat{\Sigma}(\mathbf{t}_i)^{-1} \left\{ \mathbf{y}_i - \widehat{\mu}(\mathbf{t}_i) \right\} \end{aligned}$$

for $\phi_k(\mathbf{t}_i) = (\phi_k(t_{i1}), ..., \phi_k(t_{ir_i}))^\intercal$, $\mu(\mathbf{t}_i) = (\mu(t_{i1}), ..., \mu(t_{ir_i}))^\intercal$, and $(\Sigma(\mathbf{t}_i))_{r,l} = G(t_{ir}, t_{il}) + \sigma^2 \delta_{rl}$, with $\widehat{\phi}(\mathbf{t}_i), \widehat{\mu}(\mathbf{t}_i)$, and $\widehat{\Sigma}(\mathbf{t}_i)$ being defined analogously. Then

$$\begin{aligned} \widehat{\xi}_{ik} - \widetilde{\xi}_{ik} &= \widehat{\lambda}_k \widehat{\phi}_k(\mathbf{t}_i)^\intercal \widehat{\Sigma}(\mathbf{t}_i)^{-1} \left\{ \mathbf{y}_i - \widehat{\mu}(\mathbf{t}_i) \right\} - \\ &\quad \lambda_k \phi_k(\mathbf{t}_i)^\intercal \Sigma(\mathbf{t}_i)^{-1} \left\{ \mathbf{y}_i - \mu(\mathbf{t}_i) \right\} \\ &\quad = \left\{ \widehat{\lambda}_k \widehat{\phi}_k(\mathbf{t}_i)^\intercal \widehat{\Sigma}(\mathbf{t}_i)^{-1} - \lambda_k \phi_k(\mathbf{t}_i)^\intercal \Sigma(\mathbf{t}_i)^{-1} \right\} \mathbf{y}_i + \\ &\quad \lambda_k \phi_k(\mathbf{t}_i)^\intercal \Sigma(\mathbf{t}_i)^{-1} \mu(\mathbf{t}_i) - \widehat{\lambda}_k \widehat{\phi}_k(\mathbf{t}_i)^\intercal \widehat{\Sigma}(\mathbf{t}_i)^{-1} \widehat{\mu}(\mathbf{t}_i) \\ &\quad = \int_{\mathcal{T}} \left\{ \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau})^\intercal \widehat{\Sigma}(\boldsymbol{\tau})^{-1} - \right. \\ &\quad \left. \lambda_k \phi_k(\boldsymbol{\tau})^\intercal \Sigma(\boldsymbol{\tau})^{-1} \right\} d\{ \mathbf{y}_i I_{\mathbf{t}_i \leq \boldsymbol{\tau}} \} + \\ &\qquad \int_{\mathcal{T}} \left\{ \lambda_k \phi_k(\boldsymbol{\tau})^\intercal \Sigma(\boldsymbol{\tau})^{-1} \mu(\boldsymbol{\tau}) - \right. \\ &\quad \left. \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau})^\intercal \widehat{\Sigma}(\boldsymbol{\tau})^{-1} \widehat{\mu}(\boldsymbol{\tau}) \right\} dI_{\mathbf{t}_i \leq \boldsymbol{\tau}} \end{aligned}$$

Then, Δ_{kj} can be written as

$$
\begin{aligned}
\Delta_{kj} &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big[\int_{\mathcal{T}} \Big\{ \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau}^*)^{\intercal} \widehat{\Sigma}(\boldsymbol{\tau}^*)^{-1} - \\
&\quad \lambda_k \phi_k(\boldsymbol{\tau}^*)^{\intercal} \Sigma(\boldsymbol{\tau}^*)^{-1} \Big\} d\{ \mathbf{y}_i^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \} \widehat{z}_{ij}^* \Big] + \\
&\quad n^{-\frac{1}{2}} \sum_{i=1}^{n} \Big[\int_{\mathcal{T}} \big\{ \lambda_k \phi_k(\boldsymbol{\tau}^*)^{\intercal} \Sigma(\boldsymbol{\tau}^*)^{-1} \mu(\boldsymbol{\tau}^*) - \\
&\quad \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau}^*)^{\intercal} \widehat{\Sigma}(\boldsymbol{\tau}^*)^{-1} \widehat{\mu}(\boldsymbol{\tau}^*) \Big\} d\{ I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \} \widehat{z}_{ij}^* \Big] \\
&= \int_{\mathcal{T}} \mathbf{G}_{n1}(\boldsymbol{\tau}^*) d\mathbf{H}_{n1}(\boldsymbol{\tau}^*) + \int_{\mathcal{T}} G_{n2}(\boldsymbol{\tau}^*) dH_{n2}(\boldsymbol{\tau}^*) \\
\text{where} \quad \mathbf{G}_{n1}(\boldsymbol{\tau}^*) &= \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau}^*)^{\intercal} \widehat{\Sigma}(\boldsymbol{\tau}^*)^{-1} - \lambda_k \phi_k(\boldsymbol{\tau}^*)^{\intercal} \Sigma(\boldsymbol{\tau}^*)^{-1}, \\
\mathbf{H}_{n1}(\boldsymbol{\tau}^*) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{y}_i^* \widehat{z}_{ij}^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \\
G_{n2}(\boldsymbol{\tau}^*) &= \lambda_k \phi_k(\boldsymbol{\tau}^*)^{\intercal} \Sigma(\boldsymbol{\tau}^*)^{-1} \mu(\boldsymbol{\tau}^*) - \widehat{\lambda}_k \widehat{\phi}_k(\boldsymbol{\tau}^*)^{\intercal} \widehat{\Sigma}(\boldsymbol{\tau}^*)^{-1} \widehat{\mu}(\boldsymbol{\tau}^*), \\
H_{n2}(\boldsymbol{\tau}^*) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{z}_{ij}^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*}
\end{aligned}
$$

and $I_{\boldsymbol{\delta}\mathbf{t}^* \leq \boldsymbol{\delta}\boldsymbol{\tau}^*} = I_{\{\delta_r t_r^* \leq \delta_r \tau_r^*\}_{1 \leq r \leq R}}$.

Assumption A.3 in Web Appendix A ensures that $E(\widehat{z}_{ij}^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*}) = 0$ and standard empirical process theory yields the fact that $H_{n2}(\boldsymbol{\tau}^*)$ converges weakly to a zero-mean Gaussian process under H_0 Pollard (1990). To see that $\mathbf{H}_{n1}(\boldsymbol{\tau}^*)$ similarly converges to a zero-mean Gaussian process, consider

$$
\begin{aligned}
\mathbf{H}_{n1}(\boldsymbol{\tau}^*) &= n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{y}_i^* \widehat{z}_{ij}^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \\
&= n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{y}_i^* \left[z_{ij}^* + g_j(\boldsymbol{\theta}_j, \mathbf{x}_i) - g_j(\widehat{\boldsymbol{\theta}}_j, \mathbf{x}_i) \right] I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \\
&= n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{y}_i^* \left[z_{ij}^* - \dot{g}_j(\boldsymbol{\theta}_j, \mathbf{x}_i)^{\intercal} (\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j) \right] I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} + o_p(1) \\
&= n^{-\frac{1}{2}} \sum_{i=1}^{n} z_{ij}^* \left\{ \mathbf{y}_i^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} - \right. \\
&\quad \left. E\left[\mathbf{y}_i^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \dot{g}_j(\boldsymbol{\theta}_j, \mathbf{x}_i)^{\intercal} \right] \mathcal{U}_j(\mathbf{x}_i) \right\} + o_p(1)
\end{aligned}
$$

Then, under H_0,

$$
E\left[z_{ij}^* \left\{ \mathbf{y}_i^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} - E\left[\mathbf{y}_i^* I_{\boldsymbol{\delta}_i \mathbf{t}_i^* \leq \boldsymbol{\delta}_i \boldsymbol{\tau}^*} \dot{g}_j(\boldsymbol{\theta}_j, \mathbf{x}_i)^{\intercal} \right] \mathcal{U}_j(\mathbf{x}_i) \right\} \right] = 0
$$

and $\mathbf{H}_{n1}(\boldsymbol{\tau}^*)$ converges to a zero-mean Gaussian process.

Now $\mathbf{G}_{n1} \longrightarrow_p \mathbf{0}$ and $G_{n2} \longrightarrow_p 0$ by consistency of FPCA estimates Hall et al. (2006). Therefore, by Proposition 7.27 of Kosorok (2008), $\int_{\mathcal{T}} \mathbf{G}_{nr}(\boldsymbol{\tau}^*) d\mathbf{H}_{nr'}(\boldsymbol{\tau}^*) \to 0$ as $n \to \infty$, for $r = 1, 2$. since $\mathbf{G}_{n1}, G_{n2}$ and $\mathbf{G} = \mathbf{0}$ have bounded total variation, $\int_{\mathcal{T}} \mathbf{H}_{nr}(\mathbf{t}) d\mathbf{G}_{nr}(\mathbf{t}) \rightsquigarrow \int_{\mathcal{T}} \mathbf{H}_r(\mathbf{t}) d\mathbf{G}(\mathbf{t}) = 0$ for $r = 1, 2$. Then, integrating by parts, $\Delta_{kj} = \sum_{r=1}^{2} \left[\int_{\mathcal{T}} \mathbf{G}_{nr}(\mathbf{t}) d\mathbf{H}_{nr}(\mathbf{t})\right] = \sum_{r=1}^{2} \left[\mathbf{G}_{nr}(\mathbf{t})\mathbf{H}_{nr}(\mathbf{t})|_{\mathcal{T}} - \int_{\mathcal{T}} \mathbf{H}_{nr}(\mathbf{t}) d\mathbf{G}_{nr}(\mathbf{t})\right]$
$= \sum_{r=1}^{2} \left[-\int_{\mathcal{T}} \mathbf{H}_{nr}(\mathbf{t}) d\mathbf{G}_{nr}(\mathbf{t})\right] + o_p(1) \rightsquigarrow \sum_{r=1}^{2} \left[-\int_{\mathcal{T}} \mathbf{H}(\mathbf{t}) d\mathbf{G}(\mathbf{t})\right] = 0$ because $\mathbf{G}_{nr} = o_p(1)$ and $\mathbf{H}_{nr} = O_p(1)$. The result $q_{kj} - \widetilde{q}_{kj} = o_p(1)$ follows.

To obtain the null distribution of Q, considering equations (3.10) and (3.11), observe that Q has the same asymptotic null distribution as

$$\widetilde{Q} = \widetilde{\mathbf{q}}^{\intercal}\widetilde{\mathbf{q}} = \mathbf{u}^{\intercal}\mathbf{u} + o_p(1)$$

for $\mathbf{u} = (u_{kj})_{j \in \mathcal{S}, 1 \le k \le K}$, and $u_{kj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{Q}_{ikj} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\widetilde{\xi}_{ik} - \mathbb{A}_{kj}\mathcal{U}(\mathbf{x}_i)\} z_{ij}^*$. Let $\Gamma = \mathrm{Cov}(\mathbf{u})$, and define $\Gamma^{1/2}$ to be the symmetric half matrix such that $\Gamma^{1/2}\Gamma^{1/2} = \Gamma$. Then

$$\begin{aligned}\widetilde{Q} &= \mathbf{u}^{\intercal}\Gamma^{-1/2}\Gamma\Gamma^{-1/2}\mathbf{u} + o_p(1) \\ &= \widetilde{\mathbf{u}}^{\intercal}UAU^{\intercal}\widetilde{\mathbf{u}} + o_p(1)\end{aligned}$$

where U is an orthonormal matrix of the eigenvectors of Γ, A is a diagonal matrix of the eigenvalues of Γ, and $\widetilde{\mathbf{u}} = \Gamma^{-1/2}\mathbf{u}$ is asymptotically standard multivariate normal by the central limit theorem. Noting that, because U is orthonormal, $U^{\intercal}\widetilde{\mathbf{u}}$ is also asymptotically standard normal, $\widetilde{\mathbf{u}}^{\intercal}UAU^{\intercal}\widetilde{\mathbf{u}} = \sum_{k=1}^{sK} a_k (u_k^*)^2$ where u_k^* is an element of the asymptotically standard normal $U^{\intercal}\widetilde{\mathbf{u}}$ and a_k an eigenvalue of Γ. It finally follows that $Q \sim \sum_{l=1}^{sK} a_l \chi_1^2$.

3.6.3 Justification for the form of the test statistic

In this section, we establish the relation between the score with respect to η^2, $\partial \log \mathcal{L}^*(0)/\partial(\eta^2)$, and the form used to derive the test statistic

$$E\left[\frac{\partial \log \mathcal{L}(0)}{\partial \eta} \mid \mathbb{V}\right]^2 + E\left[\frac{\partial^2 \log \mathcal{L}(0)}{\partial \eta^2} \mid \mathbb{V}\right].$$

First note that

$$\begin{aligned}\lim_{\eta\to 0}\frac{\partial \log \mathcal{L}^*(\eta)}{\partial(\eta^2)} &= \lim_{\eta\to 0}\frac{1}{2\eta\mathcal{L}^*(\eta)}\frac{\partial\mathcal{L}^*(\eta)}{\partial\eta}\\ &= \lim_{\eta\to 0}\frac{1}{2\mathcal{L}^*(\eta)}\left\{\eta^{-1}\frac{\partial\mathcal{L}^*(0)}{\partial\eta}+\frac{\partial^2\mathcal{L}^*(0)}{\partial\eta^2}+o(\eta)\right\}\\ &= \frac{1}{2\mathcal{L}^*(0)}\frac{\partial^2\mathcal{L}^*(0)}{\partial\eta^2}\end{aligned}$$

because $\partial\mathcal{L}^*(0)/\partial\eta = 0$. Now consider that

$$\begin{aligned}\mathcal{L}^*(0) &= \exp\left\{\sum_{i=1}^{n}\left(-\frac{1}{2}\left|\Sigma_{\mathbf{y}_i}\right| - \frac{1}{2}\mathbf{y}_{\mu i}^{\intercal}\Sigma_{\mathbf{y}_i}^{-1}\mathbf{y}_{\mu i}\right)\right\}\\ &= \mathcal{L}(0).\end{aligned}$$

Thus,

$$\begin{aligned}&\{\mathcal{L}^*(0)\}^{-1}\frac{\partial^2\mathcal{L}^*(0)}{\partial\eta^2}\\ &= E\left[\{\mathcal{L}(0)\}^{-1}\frac{\partial^2\mathcal{L}(0)}{\partial\eta^2}\mid\mathbb{V}\right]\\ &= E\left[\{\mathcal{L}(0)\}^{-1}\frac{\partial\mathcal{L}(0)}{\partial\eta}\mid\mathbb{V}\right]^2 - E\left[\{\mathcal{L}(0)\}^{-1}\frac{\partial\mathcal{L}(0)}{\partial\eta}\mid\mathbb{V}\right]^2 + E\left[\{\mathcal{L}(0)\}^{-1}\frac{\partial^2\mathcal{L}(0)}{\partial\eta^2}\mid\mathbb{V}\right]\\ &= E\left[\frac{\partial\log\mathcal{L}(0)}{\partial\eta}\mid\mathbb{V}\right]^2 + E\left[\frac{\partial^2\log\mathcal{L}(0)}{\partial\eta^2}\mid\mathbb{V}\right]\end{aligned}$$

And the result is established.

Bibliography

Adeyemo, A., Bentley, A. R., Meilleur, K. G., Doumatey, A. P., Chen, G., Zhou, J., Shriner, D., Huang, H., Herbert, A., Gerry, N. P. et al. (2012). Transferability and fine mapping of genome-wide associated loci for lipids in african americans. *BMC medical genetics* **13** 88.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Bentkus, V. (2005). A lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability & Its Applications* **49** 311–323.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

Castro, P., Lawton, W. and Sylvestre, E. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

Cheung, B. M. (2010). The hypertension–diabetes continuum. *Journal of cardiovascular pharmacology* **55** 333–339.

Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 405–423.

Commenges, D. and Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1** 145–156.

Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics* **8** 387–419.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M. and Crawford, D. C. (2010). Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26** 1205–1210.

Eriksson, N., Tung, J. Y., Kiefer, A. K., Hinds, D. A., Francke, U., Mountain, J. L. and Do, C. B. (2012). Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One* **7** e34442.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Gateva, V., Sandling, J. K., Hom, G., Taylor, K. E., Chung, S. A., Sun, X., Ortmann, W., Kosoy, R., Ferreira, R. C., Nordmark, G. et al. (2009). A large-scale replication study identifies tnip1, prdm1, jazf1, uhrf1bp1 and il10 as risk loci for systemic lupus erythematosus. *Nature genetics* **41** 1228–1233.

Gertheiss, J., Goldsmith, J., Crainiceanu, C. and Greven, S. (2013). Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics* **14** 447–461.

Glahn, H. R. (1968). Canonical correlation and its relationship to discriminant analysis and multiple regression. *Journal of the atmospheric sciences* **25** 23–31.

Goeman, J. J. and Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* **38** 3782–3810.

Graham, D. S. C., Morris, D. L., Bhangale, T. R., Criswell, L. A., Syvänen, A.-C., Rönnblom, L., Behrens, T. W., Graham, R. R. and Vyse, T. J. (2011). Association of ncf2, ikzf1, irf8, ifih1, and tyk2 with systemic lupus erythematosus. *PLoS genetics* **7** e1002341.

Green, J. G., McLaughlin, K. A., Berglund, P. A., Gruber, M. J., Sampson, N. A., Zaslavsky, A. M. and Kessler, R. C. (2010). Childhood adversities and adult psychiatric disorders in the national comorbidity survey replication i: associations with first onset of dsm-iv disorders. *Archives of general psychiatry* **67** 113–123.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128.

Hafler, D., Compston, A., Sawcer, S., Lander, E., Daly, M., De Jager, P., de Bakker, P., Gabriel, S., Mirel, D., Ivinson, A. et al. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. *The New England journal of medicine* **357** 851.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics* 1493–1517.

Hardoon, D., Szedmak, S. and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation* **16** 2639–2664.

Harley, J. B. et al. (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in itgam, pxk, kiaa1542 and other loci. *Nature genetics* **40** 204–210.

Holmes, M. V., Harrison, S., Talmud, P. J., Hingorani, A. D. and Humphries, S. E. (2011). Utility of genetic determinants of lipids and cardiovascular events in assessing risk. *Nature Reviews Cardiology* **8** 207–221.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 321–377.

Jiang, C. and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140** 1111–1127.

Jiang, J. (1998). Asymptotic properties of the empirical blup and blue in mixed linear models. *Statistica Sinica* **8** 861–885.

Jin, J. and Cui, H. (2010). Asymptotic distributions in the projection pursuit based canonical correlation analysis. *Science China Mathematics* **53** 485–498.

Jin, Z., Ying, Z. and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390.

Kessler, R. C., Davis, C. G. and Kendler, K. S. (1997). Childhood adversity and adult psychiatric disorder in the us national comorbidity survey. *Psychological medicine* **27** 1101–1119.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Kurreeman, F. et al. (2011). Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *The American Journal of Human Genetics* **88** 57–69.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.

Lange, C. et al. (2003). A multivariate family-based association test using generalized estimating equations: Fbat-gee. *Biostatistics* **4** 195–206.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I. et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* **62** 1120–1127.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* 673–687.

Minnier, J., Tian, L. and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106**.

Murphy, S., Rossini, A. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92** 968–976.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95** 449–465.

Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics* **39** 1–47.

Ogasawara, H. (2007). Asymptotic expansions of the distributions of estimators in canonical correlation analysis under nonnormality. *Journal of Multivariate Analysis* **98** 1726–1750.

Othus, M. and Li, Y. (2010). A gaussian copula model for multivariate survival data. *Statistics in biosciences* **2** 154–179.

Peng, J. et al. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics* **4** 53–77.

Plenge, R. M., Padyukov, L., Remmers, E. F., Purcell, S., Lee, A. T., Karlson, E. W., Wolfe, F., Kastner, D. L., Alfredsson, L., Altshuler, D. et al. (2005). Replication of putative candidate-gene associations with rheumatoid arthritis in¿ 4,000 samples from north america and sweden: Association of susceptibility with¡ i¿ ptpn22, ctla4,¡/i¿ and¡ i¿ padi4¡/i¿. *The American Journal of Human Genetics* **77** 1044–1060.

Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR.

Prevoo, M., Van't Hof, M., Kuper, H., Van Leeuwen, M., Van de Putte, L. and Van Riel, P. (1995). Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis & Rheumatism* **38** 44–48.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)* 233–243.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science* 15–32.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100** 94–108.

Romano, J. P., Wolf, M. et al. (2010). Balanced control of generalized error rates. *The Annals of Statistics* **38** 598–633.

Smoller, J., Craddock, N., Kendler, K., Lee, P., Neale, B., Nurnberger, J., Ripke, S., Santangelo, S., Sullivan, P., Buitelaar, J. et al. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis .

Solovieff, N. et al. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* .

Somers, E. C. et al. (2006). Autoimmune diseases co-occurring within individuals and within families: a systematic review. *Epidemiology* **17** 202–217.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics* 2013–2035.

Taylor, K. E. et al. (2011). Risk alleles for systemic lupus erythematosus in a large case-control collection and associations with clinical subphenotypes. *PLoS genetics* **7** e1001311.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* 47, Sage.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94** 297–311.

Turlach, B. A., Venables, W. N. and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics* **47** 349–363.

Ueda, H., Howson, J. M., Esposito, L., Heward, J., Chamberlain, G., Rainbow, D. B., Hunter, K. M., Smith, A. N., Di Genova, G., Herr, M. H. et al. (2003). Association of the t-cell regulatory gene ctla4 with susceptibility to autoimmune disease. *Nature* **423** 506–511.

Uno, H., Cai, T., Tian, L. and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**.

van Heel, D. A. et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring il2 and il21. *Nature genetics* **39** 827–829.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association* **102**.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **8** 1–27.

Wu, H. and Zhang, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association* **97** 883–897.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.

Xavier, R. J. and Rioux, J. D. (2008). Genome-wide association studies: a new window into immune-mediated diseases. *Nature Reviews Immunology* **8** 631–643.

Xia, Y. (2008). A semiparametric approach to canonical analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 519–543.

Xue-Kun Song, P. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics* **27** 305–320.

Yang, I., Fong, K., Zimmerman, P., Holgate, S. and Holloway, J. (2009). Genetic susceptibility to the respiratory effects of air pollution. *Postgraduate medical journal* **85** 428–436.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68** 49–67.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 507–564.

Zeng, D. and Lin, D. (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica* **20** 871.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94** 691–703.

Zhernakova, A., van Diemen, C. C. and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews Genetics* **10** 43–55.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871* .

Zhu, L.-X., Zhu, L.-P. and Li, X. (2007). Transformed partial least squares for multivariate data. *Statistica Sinica* **17** 1657.

Zhu, W. and Zhang, H. (2009). Why do we test multiple traits in genetic association studies? *Journal of the Korean Statistical Society* **38** 1–10.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101** 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics* **15** 265–286.